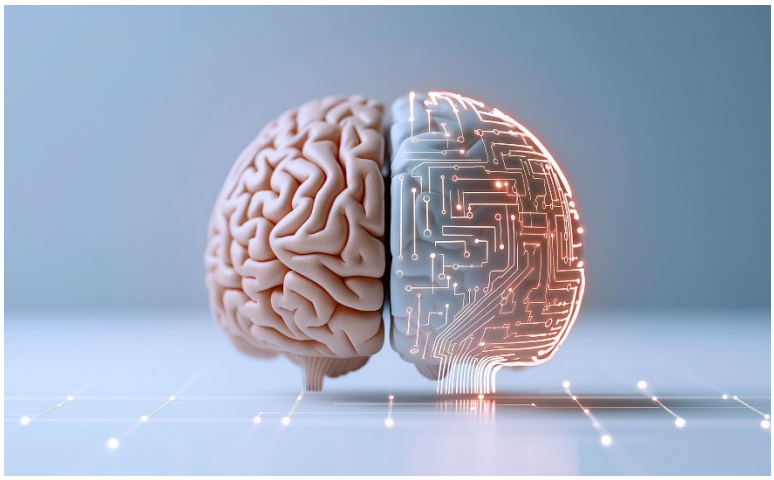
STUDY

Requested by the JURI Committee

European Parliament

Generative AI and Copyright

Training, Creation, Regulation



Policy Department for Justice, Civil Liberties and Institutional Affairs
Directorate-General for Citizens' Rights, Justice and Institutional Affairs
PE 774.095 – July 2025

EN

Generative AI and Copyright

Training, Creation, Regulation

Abstract

This study examines how generative AI challenges core principles of EU copyright law. It highlights the legal mismatch between AI training practices and current text and data mining exceptions, and the uncertain status of AI-generated content. These developments pose structural risks for the future of creativity in Europe, where a rich and diverse cultural heritage depends on the continued protection and fair remuneration of authors. The report calls for clear rules on input/output distinctions, harmonised opt-out mechanisms, transparency obligations, and equitable licensing models. To balance innovation and authors' rights, the European Parliament is expected to lead reforms that reflect the evolving realities of creativity, authorship, and machine-generated expression.

This study was commissioned by the European Parliament's Policy Department for Justice, Civil Liberties and Institutional Affairs at the request of the Committee on Legal Affairs.

This document was requested by the European Parliament's Committee on Legal Affairs.

AUTHOR

Nicola LUCCHI, PhD – Serra Hunter Professor of Comparative Law, University Pompeu Fabra, Barcelona, Spain

ADMINISTRATOR RESPONSIBLE

Mariusz MACIEJEWSKI

EDITORIAL ASSISTANTS

Ivona KLECAN, Anne DE CONINCK

LINGUISTIC VERSIONS

Original: EN

ABOUT THE EDITOR

Policy departments provide in-house and external expertise to support EP committees and other parliamentary bodies in shaping legislation and exercising democratic scrutiny over EU internal policies.

To contact the Policy Department or to subscribe for updates, please write to:
Policy Department for Justice, Civil Liberties and Institutional Affairs
European Parliament
B-1047 Brussels
Email: poldep-iust-b@europarl.europa.eu

Manuscript completed in July 2025
© European Union, 2025

This document is available on the internet at:
http://www.europarl.europa.eu/supporting-analyses

DISCLAIMER AND COPYRIGHT

The opinions expressed in this document are the sole responsibility of the authors and do not necessarily represent the official position of the European Parliament.
Reproduction and translation for non-commercial purposes are authorised, provided the source is acknowledged and the European Parliament is given prior notice and sent a copy.
© Cover image used under licence from Adobe Stock.com

CONTENTS

LIST OF ABBREVIATIONS 5

LIST OF TABLES 6

LIST OF FIGURES 7

EXECUTIVE SUMMARY 8

1. INTRODUCTION AND CONTEXT 12

1.1. Purpose and scope of the study 17

1.2. What is generative AI 18

1.3. Copyright Law in the EU: key principles 22

1.4. The challenge: copyright law and generative AI 26

2. USING COPYRIGHT-PROTECTED WORKS TO TRAIN GENERATIVE AI (INPUT SIDE) 29

2.1. Text and Data Mining (TDM) in the CDSM Directive 32

2.1.1. Understanding the technical distinction between TDM and Generative AI 37

2.1.2. Does Generative AI Training really qualify as Text and Data Mining? 41

2.1.3. Unauthorised Training and Its Legal Consequences 51

2.1.4. Beyond TDM: Structural Gaps in the CDSM Directive Framework 54

2.1.5. Anticipating the CJEU's Ruling in Case C-250/25 62

2.1.6. Comparative Jurisdictional Approaches to TDM: Lessons for EU Policy Reform 66

2.2. Implementation across Member States 73

2.3. Impact on rightsholders 76

2.4. Author's Rights and remuneration for AI training uses 78

2.4.1. Regulatory Gaps and Remuneration Challenges 80

2.5. The AI Act and transparency obligations 85

3. LEGAL STATUS OF AI-GENERATED OUTPUTS (OUTPUT SIDE) 90

3.1. Originality and authorship under EU law 91

3.1.1. Does the Human-Centric Approach Still Make Sense in the Era of Advanced Generative AI? 96

3.2. AI-assisted vs AI-generated: where to draw the line 99

3.3. Economic and Legal Challenges of AI-Generated Outputs: Disrupting Value Chains and Market Dynamics 103

3.4. Infringement and liability 106

4. POLICY OPTIONS AND RECOMMENDATIONS 110

4.0. Three-Pillar Accountability Test (orientation tool for Sections 4.1–4.6) 112

4.1. Governance and enforcement: Fragmented responsibilities 114

4.2. Improve implementations of TDM exceptions 118

4.3. Possible mechanisms for remuneration 126
4.4. Clarify protection status of AI-assisted vs AI-created works 142
4.5. Support safeguards and content traceability 148
4.6. Foster collaborative governance and legal coherence 152
4.7. Conclusion 157

REFERENCES **161**

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
AI ACT	Regulation (EU) 2024/1689 Artificial Intelligence ACT
CDSM Dir.	Directive (EU) 2019/790 on Copyright in the Digital Single Market
CJEU	Court of Justice of the European Union
DSA	Digital Services Act
EDMO	European Digital Media Observatory
EP	European Parliament
EU	European Union
EUIPO	European Union Intellectual Property Office
GPAI	General-Purpose AI
HLEG	High-Level Expert Group
InfoSoc	Directive 2001/29/EC Copyright in the Information Society Directive
IP	Intellectual Property
LLM	Large Language Model
OECD	Organisation for Economic Co-operation and Development
TDM	Text and Data Mining
RAG	Retrieval Augmented Generation

LIST OF TABLES

Table 1: Comparison of TDM Exceptions in CDSM Directive 35

Table 2: Differences between TDM and GenAI 40

Table 3: Summary box 50

Table 4: What is the AI Act doing? 89

Table 5: Copyright Eligibility of AI-Generated outputs under EU law 99

Table 6: The pillars at a glance 113

Table 7: Why the pillars matter – a quick walk-through 113

Table 8: A "traffic-light" test for draft amendments 114

Table 9: Three-Pillar Check 118

Table 10: Standardise opt-out and lawful-access conditions 120

Table 11: Three-Pillar Check 126

Table 12: Graduate menu 129

Table 13: Strengths and limits of the proposed mechanism 133

Table 14: Comparative Overview of Three Remuneration Models for AI training 137

Table 15: Differences between 2 standard software licences 139

Table 16: Three-Pillar Check 142

Table 17: Three-Pillar Check 147

Table 18: Three-Pillar Check 152

Table 19: Three-Pillar Check 156

Table 20: Scenario Outlook 2030: Strategic Futures for EU Copyright Governance 159

Table 21: Three-Pillar Check 160

LIST OF FIGURES

Figure 1: How generative AI works 20

Figure 2: Risk and Responsibility Matrix in the GenAI Copyright Context 138

EXECUTIVE SUMMARY

The integration of Generative Artificial Intelligence (GenAI) systems into creative workflows is transforming how content is processed, distributed, and accessed across the European Union. These large-scale, general-purpose computational models enable new forms of automation and synthesis, but their deployment also disrupts the established balance of rights and responsibilities within the copyright framework. While innovation is nothing new to copyright law, generative AI presents an unprecedented test of scale, opacity, and economic impact.

This study identifies five key findings:

(1) The current EU text-and-data mining (TDM) exception was not designed to accommodate the expressive and synthetic nature of generative AI training, and its application to such systems risks distorting the purpose and limits of EU copyright exceptions.

(2) Fully machine-generated outputs should remain unprotected; AI-assisted works require harmonised protection criteria.

(3) A statutory remuneration scheme is essential to bridge the growing value gap between creators and AI developers.

(4) The fragmented governance landscape underscores the need for more coherent, cross-sector institutional responses.

(5) Without timely reform, the EU risks legal uncertainty, market concentration, and cultural homogenisation.

The primary challenge today is not technological innovation, but the instrumental reinterpretation of legal principles that undermines their coherence. The proper response is not to make copyright law fit AI, but to ensure that AI development respects the core legal and policy principles of EU copyright, including authorship, originality, and fair remuneration.

Against this backdrop, this study—commissioned by the European Parliament's Committee on Legal Affairs (JURI)—examines the implications of generative AI systems for EU copyright law[1] and proposes policy options to ensure fairness, transparency, and legal clarity in the face of rapid technological change.

Copyright and Training Data: Legal Gaps and Industry Workarounds

A central focus of this study is the use of copyright-protected content as training data for generative AI systems. Article 4 of the Copyright in the Digital Single Market (CDSM) Directive provides a text-and-data-mining (TDM) exception that allows use of such content unless the rightsholder has opted out. However, the mechanism for reserving rights lacks a harmonised, machine-readable standard and

[1] Strictly speaking, "European copyright law" is a shorthand expression, as no single unified copyright system exists at the European Union level. Rather, each of the twenty-seven EU Member States retains its own national copyright legislation. The EU's role has primarily been to harmonize specific aspects of these national laws through a series of directives, resulting in a partially convergent legal framework across the Union.

presents significant scalability challenges. No current tagging protocol can reliably track duplicates or respond to evolving extraction techniques, which undermines effective implementation. In this context, the study considers whether restoring prior authorisation for generative AI training may offer a more sustainable and enforceable framework. Already, major developers are moving toward direct licensing arrangements with publishers, image banks, and other rightsholders, reflecting growing recognition of the limitations of the current exception. These developments raise important questions about legal certainty, equity, and transparency.

AI-Generated Outputs: Authorship, Protection, and Legal Uncertainty

The outputs of generative AI models challenge traditional notions of authorship and originality.[2] Under EU law, works generated entirely by machines without human intervention do not benefit from copyright protection. However, many outputs emerge from iterative human use of algorithmic tools, raising questions about authorship boundaries. Member States differ in how they interpret such hybrid authorship, leading to legal uncertainty and fragmentation across the internal market.

The study argues that clarity is urgently needed. Fully machine-generated content should remain in the public domain, while criteria for protecting AI-assisted works should be codified in EU law. The introduction of new, sui generis rights for machine-generated content is not recommended, as it risks undermining the coherence of the copyright system. In addition to legal uncertainty around hybrid authorship, AI-generated outputs resulting from automated processing raise significant economic challenges: they introduce market displacement risks, undermine traditional licensing structures, and risk concentrating value in the hands of a few dominant platforms, thereby destabilising incentives for professional creators. In addition, the study warns that moral-rights protection (attribution and integrity) is fragmented across Member States; without minimum EU alignment, authors may resort to forum-shopping to stop reputational distortions in AI outputs. The study also identifies two structural risks: the erosion of fair bargaining conditions for authors and the displacement of human creativity through automated content saturation. Both represent market failures that must be addressed to preserve a diverse, sustainable creative economy.

Fair Remuneration: Addressing the Value Gap

A key policy concern is the absence of any mechanism that ensures creators are remunerated when their works are used to train AI models. As things stand, the economic benefits generated by AI training are not currently accompanied by clear mechanisms for compensating rightsholders. This undermines the incentive structure on which copyright is based.

The study explores possible responses, including the establishment of a statutory remuneration scheme. Such a scheme could take the form of a collective licence or levy on AI outputs, administered by collective management organisations and based on transparent, auditable usage data. However,

[2] For the sake of readability, this study occasionally uses expressions such as 'AI-generated content' or 'generative outputs.' These should be understood as shorthand for 'outputs resulting from automated computational processes using AI models,' and do not imply authorship, intentionality, or agency.

such a solution would require strong safeguards, including enforceable disclosure obligations and public oversight. In parallel, the study also considers whether certain forms of AI-generated outputs—particularly where they displace human-authored content—could justify output-linked remuneration schemes as a means to preserve fair market conditions.

Governance and Enforcement: Fragmented Responsibilities

The institutional landscape for copyright and AI is currently fragmented. Responsibility is shared among national courts and authorities, collective management organisations (CMOs), the European Parliament (EP), the European Commission, the European Union Intellectual Property Office (EUIPO), and the AI Office. This diffusion of competences contributes to slow enforcement, jurisdictional gaps, and regulatory uncertainty.

In order to address immediate coordination gaps, the study recommends that the JURI Committee establish a dedicated Working Group on AI and Copyright to ensure political follow-up and structured inter-committee dialogue. In parallel, a six-month High-Level Expert Group (HLEG) could be convened to deliver enforceable technical standards and pilot remuneration prototypes—including assessing whether a machine-readable interim opt-out tag is a workable solution. Together, these two mechanisms would offer a dual track of expert input and parliamentary oversight, paving the way toward a more robust institutional framework.

For longer-term governance, the study proposes creating a specialised AI & Copyright Unit within the EU AI Office, operating in coordination with EUIPO, the European Parliament, the European Commission, and CMOs. This unit would support copyright-related audits, compliance verification, and policy alignment—ensuring legal coherence while minimising administrative costs.

A Framework for Accountability

The study proposes a 'Three-Pillar Accountability Test' to evaluate policy options, with three criteria: **epistemic accountability** (transparency about if and how copyrighted content is used in AI training), **normative accountability** (fair allocation of rights and revenues), and **systemic accountability** (effective institutional oversight). Chapter 4 maps each reform against these criteria to check its legal soundness and practical feasibility.

Policy Outlook

The study outlines a rights-centered reform pathway aimed at strengthening authorial control and enhancing legal clarity in the evolving landscape of generative AI. Among the proposed measures is the recalibration of Article 4 of the CDSM Directive, exploring a transition toward a default system of prior authorisation—supported by a unified, machine-readable permissions registry, potentially overseen by EUIPO. In parallel, developers of AI models would be expected to maintain standardised dataset logs and implement traceability tools (such as watermarking or fingerprinting), allowing for end-to-end auditing of protected content use. To address the value gap, the study proposes also a statutory remuneration mechanism that would allocate a fair share of AI-generated value to rightsholders, with compliance monitored through randomised corpus audits conducted by the EU AI Office. Additionally,

a proportionate moral rights framework would aim to safeguard authors against reputational harm. A tiered compliance structure would ensure that non-profit and open-source GPAI projects are not unduly burdened. This "yellow-label" relief (up to certain compute or revenue thresholds) would help maintain openness and innovation beyond the dominant commercial actors.

Depending on the level of regulatory action taken by the EU, this study outlines three strategic scenarios for the creative sector by 2030. In the most favourable outcome (Optimistic scenario – Guided Progress), harmonised transparency rules, enforceable remuneration mechanisms, and active EU participation in model development foster legal certainty and a thriving creative economy. A middle-ground scenario (Intermediate – Litigious Status Quo) emerges from fragmented or partial implementation, leading to legal ambiguity, uneven enforcement, and stagnant revenues. In the worst-case scenario (Regressive – Creative Erosion), continued inaction enables unchecked AI use, eroding rights, undermining creator income, and flattening cultural diversity. These scenarios illustrate what is at stake—and why timely, coordinated intervention is essential.

Conclusion

Exploring a transition toward a structured permission-based model may represent a necessary step toward restoring coherence and legal certainty within the EU copyright framework. Generative AI systems operate at a scale and opacity that EU copyright law was never designed to address. To uphold core copyright values, the EU should pursue targeted, proportionate reforms that reinforce its existing legal architecture. A phased approach could support this evolution: first, by reinforcing authors' existing rights and halting the erosion of foundational copyright principles; and then, by introducing statutory mechanisms that promote legal certainty, traceability, and fair remuneration without imposing unworkable transactional burdens.

This study outlines a path toward such reform—grounded in transparency, proportionality, and systemic coherence—so that Europe can remain both innovation-friendly and protective of creators. While there will be reasonable disagreement over the optimal regulatory path, the proposals aim to offer a balanced response that aligns technological development with cultural and legal sustainability. By reintroducing a permission-based approach, ensuring fair remuneration, and strengthening oversight, the EU can position itself as a global leader in fostering an AI-and-copyright regime that is both responsible and resilient for the future.

1. INTRODUCTION AND CONTEXT

KEY FINDINGS:

Generative AI is transforming creative workflows: Generative AI technologies are increasingly integrated into the creative workflow as computational tools, raising profound legal and ethical concerns.

AI systems are trained on datasets that include human-made works: Generative models are trained on vast datasets that often contain copyrighted material used without rightsholder consent or compensation.

Two major legal challenges arise: Whether the use of copyrighted inputs for AI training is lawful under EU law, and whether AI-generated outputs can be protected—and by whom.

EU copyright law remains human-centric: Current rules require human creativity and authorship, meaning that most AI-generated outputs fall outside the scope of protection.

Text and data mining rules are not adapted to AI: Articles 3 and 4 of the CDSM Directive were not designed for large-scale model training and do not provide legal certainty, transparency, or effective rights control.

Fragmentation and uncertainty hamper legal clarity: Divergent national implementations of the text and data mining (TDM) exceptions in the CDSM Directive across EU Member States complicate compliance and increase risk for AI developers.

Opt-Out Mechanism Is Structurally Unfit for Generative AI: Existing opt-out tools like metadata or robots.txt are ineffective for large-scale web scraping and training corpora construction.

Balancing innovation with fair remuneration is key: While the AI Act's transparency obligations may support oversight, they do not resolve core copyright challenges.

Targeted copyright reform is necessary: The study calls for clearer rules, stronger enforcement, and, where justified, carefully designed legal instruments to address the challenges posed by generative AI—without undermining core copyright principles.

Generative artificial intelligence (AI) represents a major shift in digital technology, altering how content is created and used across sectors. AI systems, such as ChatGPT[3], Gemini,[4] and Deepseek[5] — all of which are large language models (LLMs) — and DALL-E,[6] Stable Diffusion[7] and Midjourney,[8] which are

[3] See OpenAI, ChatGPT, available at: https://openai.com/chatgpt

[4] See Google DeepMind, Gemini, available at: https://deepmind.google/technologies/gemini/

[5] See Deepseek, Deepseek LLM, available at: https://deepseek.com/

[6] See OpenAI, DALL·E, available at: https://openai.com/dall-e

[7] See Stability AI, Stable Diffusion, available at: https://stability.ai/

[8] See Midjourney, Midjourney, available at: https://www.midjourney.com/home

leading generative models for images, "learn" from extensive datasets comprising diverse media, including text, images, music, and video.[9]

This shift signals a new era in which AI is increasingly integrated into creative processes, reshaping human–machine interaction.[10] Generative AI analyses large datasets to identify patterns and produce synthetic outputs that mimic original works.[11] These systems use advanced models (natural language processing, pattern recognition, etc.) to generate text in a human-like style and coherent form, based on statistical patterns learned from the training data.[12]

While generative AI opens remarkable opportunities for innovation and efficiency, it also raises significant ethical and legal challenges for intellectual property rights.[13] One concern is that generative

---

[9] It is important to stress that the language modelling task relies solely on the form of training data (surface-level patterns such as word sequences) and therefore cannot inherently lead to the learning of meaning. See, e.g., Emily M. Bender and Alexander Koller, Climbing Towards NLU: On Meaning, Form, and Understanding in the Age of Data, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020): 5185–5198, available at https://doi.org/10.18653/v1/2020.acl-main.463 (defining the term "language model" as any system trained only on the task of string prediction, whether it operates over characters, words or sentences and sequentially or not).

[10] For a discussion of the evolving dynamics of human-machine collaboration, see, e.g. See e.g. Minglun Ren et al., H. Human-machine Collaborative Decision-making: An Evolutionary Roadmap Based on Cognitive Intelligence, 15 Int J of Soc Robotics 1101–1114 (2023); Tony McCaffrey and Lee Spector, An approach to human–machine collaboration in innovation. 32 AI EDAM, 1-15 (2018); Hyunjin Kang and Chen Lou, AI agency vs. human agency: understanding human–AI interactions on TikTok and their implications for user engagement, 27 Journal of Computer-Mediated Communication 1-13 (2022); Francesco Semeraro, et al., Human–robot collaboration and machine learning: A systematic review of recent research, 79 Robotics and Computer-Integrated Manufacturing, 1-16 (2023); Liana Razmerita et al., Collaboration in the Machine Age: Trustworthy Human-AI Collaboration. In: Virvou, M., Tsihrintzis, G.A., Jain, L.C. (eds) Advances in Selected Artificial Intelligence Areas. Learning and Analytics in Intelligent Systems, Springer (2020); Jean-Michel Hoc, From human-machine interaction to human-machine cooperation, 43 Hergonomics 833, 843 (2000).

[11] OECD, OECD Framework for the Classification of AI Systems, OECD Digital Economy Papers, No. 323, OECD Publishing, Paris (2022), at 45–46, available at https://doi.org/10.1787/cb6d9eca-en (defining generative models as involving the discovery and learning of the patterns and distribution of input data, enabling the generation of new plausible examples that could be part of the original distribution). See also Ian J. Goodfellow et al., Generative Adversarial Nets. 2 Proceedings of the 27th International Conference on Neural Information Processing Systems 2672-2680 at 2672 (2014) (describing generative models as capturing the data distribution and generating new samples by transforming random noise).

[12] See Rishi Bommasani et al., On the Opportunities and Risks of Foundation Models, Preprint at https://doi.org/10.48550/arXiv.2108.07258 (2022), at 48/49 (noting that foundation models are trained via self-supervision to learn co-occurrence patterns in data sequences, which enables them to generate fluent, human-like outputs based on statistical regularities rather than explicit understanding); Luciano Floridi, AI as Agency Without Intelligence: On ChatGPT, Large Language Models, and Other Generative Models, 36 Philosophy & Technology 1-7 at 2(2023), (noting that large language models process the formal structure of texts statistically, enabling them to generate outputs that imitate semantic coherence without actual understanding).

[13] While not exhaustive, the following sources offer extensive insights and diverse perspectives on the complex issues surrounding generative AI and intellectual property: Amir Khoury, "Intellectual Property Rights for Hubots: On the Legal Implications of Human-like Robots as Innovators and Creators," (2017) 35 Cardozo Arts & Entertainment Law Journal 635; Enrico Bonadio et al., "Intellectual Property Aspects of Robotics," (2018) 9 European Journal of Risk Regulation 655; Mark Lemley and Bryan Casey, "Remedies for Robots," (2019) 86 University of Chicago Law Review 1311; Enrico Bonadio and Luke McDonagh, "Artificial Intelligence as Producer and Consumer of Copyright Works: Evaluating the Consequences of Algorithmic Creativity," (2020) Intellectual Property Quarterly 112; Ryan Abbott, The Reasonable Robot (Cambridge, Cambridge University Press, 2020); Tim Dornis, Artificial Creativity: Emergent Works and the Void in Current Copyright Doctrine, 22 Yale Journal of Law & Technology 1 (2020); Giuseppe Abbamonte, The Rise of the Artificial Artist: AI Creativity, Copyright and Database Right, 43 European Intellectual Property Review 702 (2021); Jenny Quang, Does Training AI Violate Copyright Law? 36 Berkeley Technology Law Journal 1407 (2021); Benjamin Sobel, A Taxonomy of

AI is "eating the creativity of the world" – using large amounts of human-created content without compensating the creators.[14] Philosophically, copyright and related rights emphasize the fundamental need to remunerate human authors.[15] Given that generative AI systems achieve their impressive capabilities precisely by analysing and learning from existing human creations, it is essential to address this issue. The "parasitic usurpation of the market for literary and artistic productions" by generative AI suggests that original human authors deserve fair compensation for their contributions to these AI models.[16] Otherwise, this undermines the incentive structure on which copyright is based.[17]

This is not the first time copyright has faced a technological challenge. From the printing press to photography to digital media, copyright law has evolved without abandoning its foundations. The current situation is not unprecedented—history, in many ways, is repeating itself.[18] What makes the

Training Data: Disentangling the Mismatched Rights, Remedies, and Rationales for Restricting Machine Learning, in R. Hilty et al. (eds.), Artificial Intelligence and Intellectual Property (Oxford, Oxford University Press, 2021) 221-242; Mark Lemley and Bryan Casey, Fair Learning, 99 Texas Law Review 743 (2021); Ruth Taplin, Artificial Intelligence, Intellectual Property, Cyber Risk and Robotics (Routledge, 2022); Enrico Bonadio et al., Can Artificial Intelligence Infringe Copyright? Some Reflections, in R. Abbott (ed.), Research Handbook on Intellectual Property and Artificial Intelligence (Cheltenham, Edward Elgar, 2022); Giorgio Franceschelli and Mirco Musolesi, Copyright in Generative Deep Learning, (2022) 4 Data & Policy e17; Jan Smits and Tijn Borghuis, Generative AI and Intellectual Property Rights, in B. Custers and E. Fosch-Villaronga (eds.), Law and Artificial Intelligence, Information Technology and Law Series, vol. 35, T.M.C. Asser Press, The Hague (2022); Martin Kretschmer et al., Artificial Intelligence and Intellectual Property: Copyright and Patents—A Response by the CREATe Centre to the UK Intellectual Property Office's Open Consultation, 17 Journal of Intellectual Property Law & Practice 321-326 (2022); Gil Appel et al., Generative AI Has an Intellectual Property Problem, Harvard Business Review, 7 April 2023, available at: https://hbr.org/2023/04/generative-ai-has-an-intellectual-property-problem; Alain Strowel, ChatGPT and Generative AI Tools: Theft of Intellectual Labor?" 54 International Review of Intellectual Property and Competition Law 491(2023); Peter Georg Picht and Florent Thouvenin, AI and IP: Theory to Policy and Back Again – Policy and Research Recommendations at the Intersection of Artificial Intelligence and Intellectual Property, 54 International Review of Intellectual Property and Competition Law 916–940 (2023); Christophe Geiger, Elaborating a Human Rights Friendly Copyright Framework for Generative AI, 55 International Review of Intellectual Property and Competition Law 1129–1165 (2024); Nicola Lucchi, ChatGPT: A Case Study on Copyright Challenges for Generative Artificial Intelligence Systems, 15 European Journal of Risks Regulation 602-624 (2024).

14 See Kalpana Tyagi, "Copyright, Text & Data Mining and the Innovation Dimension of Generative AI," (2024) 19 J. Intell. Prop. L. & Prac. 557, 567 (reflecting on concerns that generative AI models appropriate human intellectual output at scale, and citing Marc Andreessen's statement that "software is eating the world").

15 See, e.g., Georg Friedrich Wilhelm Hegel, Hegel's Philosophy of Right § 69 (Thomas M. Knox trans., Clarendon Press 1967) (1821); Immanuel Kant, Kritik der Urteilskraft, in Kant's Kritik of Judgment § 46 (J.H. Bernard trans., Macmillan and Co. 1892) (1790); John Locke, Two Treatises of Government, in The Works of John Locke § 27 (1727); see also Justin Hughes, The Philosophy of Intellectual Property, 77 Geo. L.J. 287 (1988).

16 See Martin Senftleben, Martin, AI Act and Author Remuneration - A Model for Other Regions? (February 24, 2024) at 3. Available at SSRN: https://ssrn.com/abstract=4740268.

17 See e.g. S. Alex Yang and Angela Huyue Zhang, Generative AI and Copyright: A Dynamic Perspective (February 4, 2024). Available at SSRN: https://ssrn.com/abstract=4716233; David De Cremer et al., How Generative AI Could Disrupt Creative Work, Harvard Bus. Rev. (Apr. 13, 2023). Available https://hbr-org.sare.upf.edu/2023/04/how-generative-ai-could-disrupt-creative-work; See Martin Senftleben, Generative AI and Author Remuneration, 54 IIC – International Review of Intellectual Property and Competition Law 1535 (2023); Frank Pasquale and Haochen Sun, Consent and Compensation: Resolving Generative AI's Copyright Crisis," 110 Virginia Law Review Online 207–47 (2024); U.S. Copyright Office, Copyright and Artificial Intelligence, Part 3: Generative AI Training (Pre-Publication Version, May 2025) at 48. Available at https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf

18 See e.g. Justin Hughes, A Short History of "Intellectual Property" in Relation to Copyright, 33 Cardozo L. Rev. 1293, 1323 (2012) (noting—already in the context of earlier technological debates—that many contemporary copyright controversies

generative AI debate distinct is not the pace or scale of innovation, but the risk that legal interpretations—both in terms of how content is ingested and how outputs are treated—may distort rather than evolve the system. The core challenge is not to reinvent copyright, but to preserve its integrity through principled evolution.

Consequently, two crucial legal questions emerge within the existing European Union (EU) copyright framework:

Input Side: Is the utilization of copyrighted works for training AI models legally permissible under EU law, and if so, under what specific conditions?

Output Side: Can outputs generated by AI systems qualify for copyright protection, and who, if anyone, holds the rights to such content? What legal and economic mechanisms are necessary to ensure fair attribution and remuneration in light of the structural impact of AI-generated content on creative markets?

The current EU copyright legal structure, originally designed around human authorship, provides authors with exclusive rights including reproduction, distribution, and adaptation.[19] The training process of generative AI inherently involves the reproduction of extensive amounts of copyrighted material into training datasets, thereby engaging the exclusive right of reproduction under EU copyright law.[20] The CDSM Directive introduces limited exceptions for text and data mining (TDM) under Articles 3 and 4, attempting to facilitate lawful use of data. However, ambiguities around conditions such as the rightsholders opt-out provisions leave significant uncertainty regarding the applicability of these exceptions.

Moreover, EU copyright law traditionally rests on the criterion of originality as established by the Court of Justice of the European Union (CJEU), requiring an author's personal intellectual creation and human creative input.[21] AI-generated outputs challenge this criterion due to their algorithmic nature and absence of direct human authorship, complicating their qualification for copyright protection.

The legal complexity is further intensified by the EU's partially harmonized copyright landscape, resulting in varying interpretations and enforcement practices across Member States. Such

reflect "just a little bit of history repeating); Brad Sherman and Leanne Wiseman (eds), Copyright and the Challenge of the New (Kluwer Law International, 2012), at 1, (observing that "one of the most challenging things about copyright law is that it is constantly subject to change.")

[19] Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society, OJ L 167, 22.6.2001, pp. 10–19, Art. 2 (Reproduction right), Art. 3 (Communication to the public), and Art. 4 (Distribution right).

[20] See EUIPO, The Development of Generative Artificial Intelligence from a Copyright Perspective (2025). Available at https://www.euipo.europa.eu/en/news/euipo-releases-study-on-generative-artificial-intelligence-and-copyright at 154–155 (noting that LLMs and image generation models can memorise and regurgitate long sequences or images from training data, including potentially copyright-protected content, particularly when original or unique); See U.S. Copyright Office, Copyright and Artificial Intelligence, Part 3: Generative AI Training (Pre-Publication Version, May 2025) at 27 et seq. Available at https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf (noting that training generative models typically requires making copies of the training examples… which implicate the reproduction right when those examples are protected by copyright).

[21] See e.g. C-5/08 Infopaq [2009] ECLI:EU:C:2009:465; C-393/09 BSA [2010] ECLI:EU:C:2010:816; C-145/10 Painer [2011] ECLI:EU:C:2011:798.

fragmentation poses significant challenges for cross-border generative AI systems, undermining legal certainty and effective compliance.

Recent legal and policy developments have amplified the urgency of resolving these challenges. High-profile disputes involving artists and copyright holders against platforms such as OpenAI highlight practical concerns about unauthorized uses of copyrighted works in AI training datasets.[22] Concurrently, the EU Artificial Intelligence Act (AI Act)[23] introduces another regulatory layer specifically addressing AI systems, which further complicates the already intricate intersection with copyright.

The core tension in this debate is between those promoting open AI innovation and creative industries worried about unauthorized, unpaid use of their work. Any policy solution must balance innovation with protecting creators' rights to fair compensation and credit. In practice, regulators should weigh qualitative impacts alongside quantitative, financial outcomes.

This study, requested by the JURI Committee, aims to provide clarity on these complex issues. It examines the technological underpinnings, explores legal nuances within the existing EU copyright framework, and assesses ongoing policy debates. The subsequent chapters will delve into these dimensions more deeply, offering policy recommendations that holistically consider the EU's overarching regulatory goals, including creativity, innovation, consumer protection, digital transformation, and economic competitiveness.

Clarification on Scope: Use of the Term "Generative AI"

Throughout this study, the term generative AI refers primarily to general-purpose AI (GPAI) models designed to compute outputs across multiple modalities—such as text, images, music, or code—based on large-scale training datasets. These include, but are not limited to, large language models (LLMs) and image generation models. In line with Recital 105 and Article 53 of the EU Artificial Intelligence Act, the analysis focuses on the copyright implications of these GPAI systems, particularly with respect to the use of protected content for training and the legal status of AI-generated outputs.

---

[22] There are numerous pending lawsuits involving companies that develop and deploy generative AI technologies, particularly in the United States. See, e.g., The New York Times Co. v. Microsoft Corp. & OpenAI, Inc., No. 1:23-cv-11195 (S.D.N.Y. filed Apr. 4, 2025); Andersen v. Stability AI Ltd., No. 3:23-cv-00201 (N.D. Cal. filed Jan. 13, 2023) (alleging unauthorized use of copyrighted artworks in AI training datasets). Similar disputes have emerged in Europe. See, e.g., Getty Images (US), Inc. v. Stability AI Ltd., [2023] EWHC (Ch) 3090 (UK High Court); Union Nationale des Éditeurs et Auteurs v. Meta Platforms, Inc., Paris Judicial Court (filed Apr. 2025), available at https://www.sne.fr/actu/unis-auteurs-et-editeurs-assignent-meta-pour-imposer-le-respect-du-droit-dauteur-aux-developpeurs-doutils-dintelligence-artificielle-generative; see also GEMA, GEMA Files Model Action to Clarify AI Providers' Remuneration Obligations in Europe, GEMA (Apr. 17, 2024), available at https://www.gema.de/en/w/gema-files-lawsuit-against-openai. In parallel, broader industry conflicts have also emerged: in 2023, the Hollywood screenwriters' strike prominently featured demands to restrict unregulated use of generative AI in scriptwriting, reflecting deep tensions between creators and platform providers over authorship, attribution, and compensation. See e.g. Molly Kinder, Hollywood writers went on strike to protect their livelihoods from generative AI. Their remarkable victory matters for all workers, (April 12, 2024). Available at https://www.brookings.edu/articles/hollywood-writers-went-on-strike-to-protect-their-livelihoods-from-generative-ai-their-remarkable-victory-matters-for-all-workers/.

[23] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139, and (EU) 2019/2144, and Directives 2014/90/EU, (EU) 2016/797, and (EU) 2020/1828 (Artificial Intelligence Act), OJ L 1689, 12.7.2024, p. 1–144 (hereinafter: EU AI ACT).

1.1. Purpose and scope of the study

This study is prompted by the growing urgency to address the legal ambiguities and structural tensions that generative artificial intelligence (GenAI) poses to the European Union's copyright framework. As AI technologies increasingly permeate creative and productive sectors, it becomes necessary to evaluate whether current uses of AI comply with the legal architecture, and where necessary, determine if adjustments are warranted to preserve the integrity of copyright principles. At its core, the study responds to a twofold concern: on the one hand, the need to assess whether the widespread use of protected works in AI training complies with EU law and to determine whether legal clarification or reform is needed; on the other, the need to safeguard technological innovation and promote responsible AI development within a competitive European digital economy.

The dual aim of this research is to propose clear, pragmatic, and legally sound policy options that strike a fair balance between the rights and interests of human creators, and the innovation potential of developers and users of generative AI. It is neither a call for deregulation nor a defence of the status quo. Rather, it reflects the necessity of a targeted legal and policy response that recognises the uniqueness of AI technologies—particularly general-purpose models—and their far-reaching implications for authorship, ownership, and remuneration in the creative economy. In particular, the study does not presume that EU copyright law must be adapted to accommodate AI systems. Instead, it proceeds in two analytical steps: first, assessing whether current uses of generative AI systems comply with EU copyright provisions; and second, identifying where violations or regulatory gaps exist that may justify proportionate legislative or enforcement responses.

Building on this two-step framework, the study focuses on two interconnected dimensions of the copyright-AI nexus: the use of protected works as input during the training phase of generative models, and the output obtained through automated processes by these systems. The analysis is premised on the understanding that both dimensions raise complex legal questions, some of which lie at the intersection of existing copyright provisions, emerging AI regulation, and fundamental principles of intellectual property law—especially those related to originality, human authorship, and fair compensation.

On the input side, the core question is whether current exceptions—particularly those related to text and data mining (TDM) introduced by Articles 3 and 4 of the CDSM Directive—can meaningfully accommodate the scale and nature of generative AI training. The study will explore the practical and legal limits of the TDM exceptions, the role of opt-out mechanisms, and the challenges raised by the non-transparent or open-ended nature of many training datasets. At the same time, the study acknowledges the transparency obligations introduced by the AI Act as a promising, though still embryonic, regulatory response. It will analytically assess how these obligations intersect with copyright law and how they might be implemented in practice.

On the output side, the analysis will interrogate the extent to which AI-generated content—particularly when produced without substantial human intervention—can or should benefit from copyright protection. This aspect implicates long-standing jurisprudence by the Court of Justice of the European Union (CJEU) on originality and intellectual creation, but it also raises new questions regarding the

status of derivative works, the potential for market substitution, and the legitimacy of granting rights (or related rights) in non-human outputs. The study will not propose artificial or speculative categories but will evaluate existing legal tools and explore whether complementary or sui generis rights may be justified under specific conditions. At the same time, it will examine how such outputs impact creative markets, and what legal and economic mechanisms might be necessary to safeguard fair attribution and remuneration for human authors.

More broadly, the study will also explore whether and how authors and rightsholders should be remunerated when their works are used in training AI systems. This inquiry builds on the recognition that human creativity is not merely a raw material to be mined but a legal and cultural resource that underpins Europe's creative sectors. In this regard, the study does not approach copyright as a barrier to innovation, but rather as a foundational mechanism for sustaining it— one that may require stronger enforcement, and in limited cases, reform, to respond to technological circumvention.

The study therefore aims to inform the European Parliament, and specifically the JURI Committee, by delivering not just a descriptive legal analysis, but a forward-looking framework for reform. It will incorporate technological and market developments, legal doctrine, policy considerations, and stakeholder perspectives. Where appropriate, it will reference comparative examples and international standards, while remaining anchored in the EU's legal and institutional context. The objective is not to advocate for one definitive solution, but to outline a set of coherent policy paths through which the EU can ensure both the protection of creative works and the responsible evolution of generative AI.

This study is based exclusively on desk research. Due to the short timeline of this study, no new empirical consultations were conducted. However, to reinforce the policy recommendations with existing stakeholder insights, the report integrates and synthesises publicly available position papers, industry statements, and consultation responses submitted to the European Commission and Member States in the context of the AI Act, the CDSM Directive implementation, and recent copyright consultations. This ensures that the study remains grounded in a representative set of concerns already expressed across the creative, technological, and legal sectors. The study also benefited from informal exchanges with technology experts, stakeholders from the creative and AI sectors, academic researchers, collective management bodies and legal scholars working in the field of intellectual property and digital regulation. While these discussions did not form part of a formal consultation process, they offered valuable insights that helped shape the legal analysis and inform the policy options proposed.

1.2. What is generative AI

Generative artificial intelligence (AI) refers to a subcategory of AI systems capable of computing or assembling synthetic content based on input data—such as text, images, audio, or video—that mimics human creativity.[24] According to Article 3(1) of the EU Artificial Intelligence Act, an 'AI system' is

[24] See OECD, *OECD Framework for the Classification of AI Systems*, OECD Digital Economy Papers, No. 323, OECD Publishing, Paris (2022), at 45, available at https://doi.org/10.1787/cb6d9eca-en (describing generative AI as involving the identification and internalization of patterns and distributions in input data, enabling the creation of novel yet statistically plausible outputs that resemble the original data); AI Act, recital 99 (providing that large generative AI models

defined as a machine-based system designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, inferring from the input it receives how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.[25] Recital 105 of the AI Act further clarifies that general-purpose generative models present both innovation potential and serious challenges for artists, authors, and other creators, as their development relies on large-scale access to data—much of which may be protected by copyright.[26] While Recital 105 of the EU AI Act highlights the legal and economic challenges posed by generative models, Recital 99 clarifies their regulatory classification. The scope of this study treats generative AI as a technologically and legally significant subset of general-purpose AI (GPAI). This view reflects Recital 99 of the AI Act, which states that "large generative AI models are a typical example of a general-purpose AI model." This classification is reinforced by a 2025 report for the European Economic and Social Committee, which explains that foundation models are often referred to in policy contexts as general-purpose AI systems due to their broad applicability across domains and tasks.[27] Similarly, guidance from UNESCO highlights that many generative AI tools—particularly large language models—are built on general-purpose transformer architectures, reinforcing the view that generative capabilities typically emerge from foundational infrastructures.[28]

Similarly, the European Commission describes AI more broadly as "systems that display intelligent behaviour by analysing their environment and taking actions—with some degree of autonomy—to achieve specific goals".[29] Generative models—including large language models (LLMs) and diffusion models—are trained on extensive datasets and operate by identifying complex patterns in the data,

---

are a paradigmatic example of general-purpose AI, as they can flexibly generate diverse content—such as text, audio, images, or video—suitable for a wide range of tasks); Artificial Intelligence Study: Notice of Inquiry, 88 Fed. Reg. 59942, 59948–49 (Aug. 30, 2023) (defining "generative AI" as AI applications that generate outputs in the form of expressive material, including text, images, audio, or video).

[25] See art. 3(1) AI Act.

[26] According to Recital 105 of the AI Act, general-purpose AI models "in particular large generative AI models, capable of generating text, images, and other content, present unique innovation opportunities but also challenges to artists, authors, and other creators and the way their creative content is created, distributed, used and consumed. The development and training of such models require access to vast amounts of text, images, videos and other data. Text and data mining techniques may be used extensively in this context for the retrieval and analysis of such content, which may be protected by copyright and related rights. Any use of copyright protected content requires the authorisation of the rightsholder concerned unless relevant copyright exceptions and limitations apply. Directive (EU) 2019/790 introduced exceptions and limitations allowing reproductions and extractions of works or other subject matter, for the purpose of text and data mining, under certain conditions. Under these rules, rightsholders may choose to reserve their rights over their works or other subject matter to prevent text and data mining, unless this is done for the purposes of scientific research. Where the rights to opt out has been expressly reserved in an appropriate manner, providers of general-purpose AI models need to obtain an authorisation from rightsholders if they want to carry out text and data mining over such works".

[27] See European Economic and Social Committee, Generative AI and foundation models in the EU – Uptake, opportunities, challenges, and a way forward, Publications Office of the European Union, 2025. Available at https://data.europa.eu/doi/10.2864/8377116

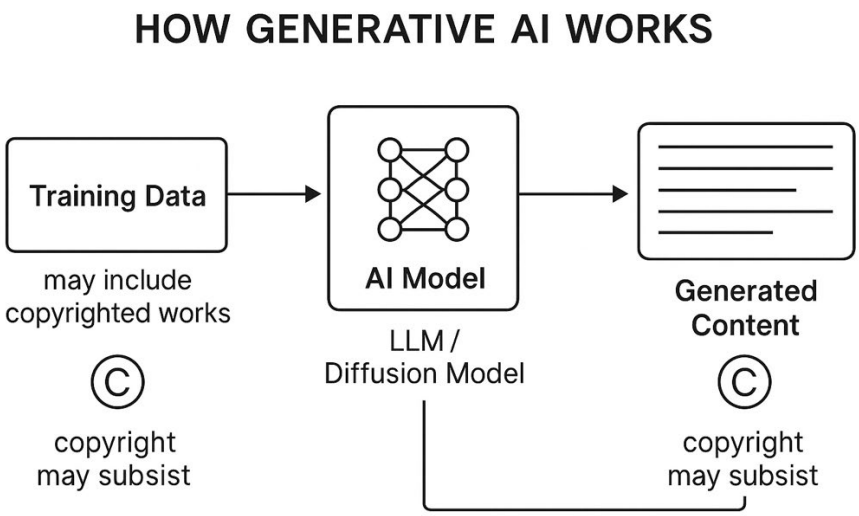
[28] See OECD, OECD Framework for the Classification of AI systems, OECD Digital Economy Papers, No. 323, OECD Publishing, Paris (2022). Available at https://doi.org/10.1787/cb6d9eca-en.

[29] See European Commission, Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions: Artificial Intelligence for Europe, COM(2018) 237 final, Brussels, 25 April 2018.

enabling them to produce outputs that often closely resemble those of human origin. The foundational architecture of these models involves training on billions of data points, often scraped from the internet, including web content, literature, visual art, and audio recordings. As acknowledged in Recital 105 of the AI Act, this process frequently relies on text and data mining techniques to retrieve and analyse material—much of which is subject to copyright and related rights protection.[30] EU law requires authorisation from rightsholders for such uses, unless relevant exceptions apply. While Directive (EU) 2019/790 introduces exceptions and limitations for text and data mining, their applicability to AI training remains uncertain. Moreover, rightsholders may reserve their rights through an opt-out mechanism, further contributing to a complex and contested legal environment for AI developers.[31]

The figure below illustrates the standard process by which generative AI operates:

Figure 1: How generative AI works



Training Data: Massive datasets are compiled, often scraped from online sources. These may include copyrighted works such as literature, photography, music, and academic publications. Copyright may subsist in these inputs.

AI Model: The data is used to train the AI system—typically an LLM or a diffusion model—allowing it to 'learn' patterns and structures without understanding content in a human sense.

Generated Content: The trained model produces outputs that may resemble human-authored content. Copyright may or may not subsist in the output, depending on human input and national legal interpretations.

[30] See supra note 26.

[31] See e.g. Eleonora Rosati, Copyright and the CDSM Directive: A Commentary, Oxford University Press, (2021) at 60 (discussing Article 4 in detail and highlights the legal uncertainty around the opt-out, especially in the context of large-scale TDM and AI training).

Technically, these systems rely on natural language processing (NLP), pattern recognition, and probabilistic modelling to synthesise seemingly coherent results.[32] However, as highlighted in recent scholarly work, generative AI performs its functions "acting without human understanding".[33] These systems can replicate linguistic and aesthetic structures, but they lack consciousness, intentionality, or the ability to comprehend meaning.[34] As a result, they do not "learn" like humans do. Whereas human learning integrates meaning, reflection, and contextual knowledge, AI models operate by extracting and reproducing statistical patterns from mined materials—effectively copying fragments of existing works rather than understanding them.

This distinction between human creativity and machine output is crucial. Human authors imbue their works with personal expression, cultural context, and intention—elements grounded in human subjectivity and personhood. In contrast, AI-processed works result from statistical pattern recognition and lack the legal hallmark of original intellectual creation. This cognitive gap has profound legal implications. Human learners can restate an idea in a novel way without infringing copyright, thanks to the idea/expression dichotomy. In contrast, AI systems must ingest, copy, and computationally process the actual expressions of protected works in order to produce outputs. As such, even where no recognisable similarity exists between the training data and the final output, this does not alter the legal characterisation of the training process itself—as one involving protected acts of reproduction. This epistemic and ontological divide not only informs the legal analysis of the training process but also underpins the current exclusion of AI outputs from authorship under copyright law, regardless of future technological developments. [35]

At a broader level, generative AI poses broader systemic challenges. As the technology evolves toward autonomous agents capable of multi-modal interaction,[36] the line between human and artificial creation becomes increasingly blurred.[37] This raises not only legal and economic concerns but also ethical ones.

---

[32] See e.g. Emily M. Bender and Alexander Koller, Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data, in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (Association for Computational Linguistics, Online, 2020) pp 5185–98 (defining the term "language model" as any system trained only on the task of string prediction, whether it operates over characters, words or sentences and sequentially or not); Yoav Goldberg, Neural Network Methods for Natural Language Processing (Cham, Springer 2017) at. 105.

[33] See Luciano Floridi, AI as Agency without Intelligence: on Chat GPT, Large Language Models and Other Generative models 36 Philosophy & Technology 1, 6 (2023) (defining this as "agere sine intelligere").

[34] Ibidem.

[35] See U.S. Copyright Office, Copyright and Artificial Intelligence, Part 3: Generative AI Training (Pre-Publication Version, May 2025) at 48. Available at https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf (noting that "Humans retain only imperfect impressions of the works they have experienced, filtered through their own unique personalities, histories, memories, and worldviews. Generative AI training involves the creation of perfect copies with the ability to analyze works nearly instantaneously.")

[36] "Autonomous agents" here refers to AI systems that can operate independently, often integrating multiple input/output modalities (e.g., voice, image, and text), and perform actions across digital environments. Examples include virtual assistants capable of planning a trip based on spoken commands, generating images, and booking tickets online.

[37] See e.g. Zane Durante et al., Agent AI: Surveying the Horizons of Multimodal Interaction (arXiv:2401.03568v2) (2024). Available at https://arxiv.org/abs/2401.03568; World Economic Forum, Navigating the AI Frontier: A Primer on the Evolution and Impact of AI Agents, December 2024, https://www.weforum.org/publications/navigating-the-ai-frontier-a-primer-on-the-evolution-and-impact-of-ai-agents (arguing that the evolution of AI agents into autonomous,

What happens when human creators are being outcompeted by machines trained on their own works — without consent, without remuneration, or any opportunity to contest the process?

1.3. Copyright Law in the EU: key principles

Copyright law in the European Union is grounded in a set of key principles designed to promote creativity, protect authors' rights, and ensure access to culture and knowledge. One of the most fundamental requirements for copyright protection in the EU is that a work must be *original*.[38] According to case law from the Court of Justice of the European Union (CJEU), a work is original if it is the result of the author's own intellectual creation.[39] This implies that there must be identifiable human involvement and creative choices in the making of the work. Consequently, works that are generated entirely by artificial intelligence (AI) systems, without meaningful human input, typically do not qualify for protection under current EU copyright rules.

Copyright arises automatically and gives rightsholders a broad set of exclusive rights—including the right to reproduce, distribute, communicate, and adapt their works.[40] These rights are balanced by a number of exceptions and limitations that are designed to serve the public interest, including in areas such as education, research, and criticism.[41] As consistently held by the CJEU, exceptions and limitations must be interpreted strictly in line with Article 5 of the InfoSoc Directive,[42] though such interpretation must also respect their underlying purpose and ensure a fair balance with fundamental

multimodal systems is ushering in a new era of human–machine collaboration where agents "plan, learn and make decisions based on a comprehensive understanding of their environment and user needs").

38 While the Berne Convention does not expressly require that works be "original" to qualify for copyright protection, most national laws have incorporated such a requirement. For a comparative analysis, see Elizabeth F. Judge and Daniel Gervais, Of Silos and Constellations: Comparing Notions of Originality in Copyright Law, 27 Cardozo Arts & Entertainment Law Journal 375, 399 (2009).

39 See e.g. C-05/08, Infopaq International v. Danske Dagblades Forening (2009) ECLI:EU:C:2009:465 (Infopaq) (setting out the EU originality standard for copyright protection); C-145/10, Eva-Maria Painer v Standard VerlagsGmbH and Others, ECLI:EU:C:2011:798; C-604/10, Football Dataco Ltd and Others, ECLI:EU:C:2012:115;Case C-310/17, Levola Hengelo BV v. Smile Foods BV, ECLI:EU:C:2018:899.

40 See Berne Convention for the Protection of Literary and Artistic Works, Sept. 9, 1886, as amended Sept. 28, 1979, S. Treaty Doc. No. 99-27 (1986), 1161 U.N.T.S. 3 (particularly Articles 5(2), 9, 11, 11bis, and 12, which establish the automatic nature of copyright and the exclusive rights of reproduction, communication, and adaptation).

41 See art. 5 of the Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the Harmonisation of Certain Aspects of Copyright and Related Rights in the Information Society, OJ L 167/10

42 Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the Harmonisation of Certain Aspects of Copyright and Related Rights in the Information Society, 2001 O.J. (L 167), 10–19 (hereinafter: InfoSoc Directive).

rights.[43] In the digital age, this balance has become more difficult to achieve, especially as emerging technologies like generative AI challenge traditional legal concepts such as authorship and originality.[44]

One of the most significant recent developments in EU copyright law is the adoption of the *Copyright in the Digital Single Market Directive* (Directive (EU) 2019/790),[45] which introduced two exceptions for *text and data mining* (TDM). Article 3 allows research organisations and cultural heritage institutions to carry out TDM for scientific research purposes. Article 4, more relevant to the AI context, provides a broader exception that permits TDM by anyone—provided the rightsholders has not expressly reserved their rights in an appropriate manner, for instance by using machine-readable means.

Although this opt-out mechanism was intended to give rightholders control over reuse of their content, it introduces substantial complexity and will likely render the Article 4 exception unworkable in practice (as many scholars have noted).[46] This is particularly true in the context of generative AI training, which relies on large and diverse datasets that typically include protected content. If rightsholders exercise the opt-out widely—something that is not only legally permitted but practically encouraged—it may result in incomplete or biased training datasets, undermining both the performance and reliability of AI systems. Moreover, the lack of clear and harmonised standards for expressing the opt-out may

[43] See e.g. Case C-348/87, Stichting Uitvoering Financiële Acties v. Staatssecretaris van Financiën, ECLI:EU:C:1989:246, para. 13; Case C-476/01, Kapper, ECLI:EU:C:2004:261, para. 72; and Case C-36/05, Commission v. Spain, ECLI:EU:C:2006:672, para. 31; Case C-5/08, Infopaq International A/S v. Danske Dagblades Forening, ECLI:EU:C:2009:465, para. 56; Case C-277/10, Martin Luksan v. Petrus van der Let, ECLI:EU:C:2012:65, para. 101; Case C-138/16, Staatlich genehmigte Gesellschaft der Autoren, Komponisten und Musikverleger registrierte Genossenschaft mbH (AKM) v. Zürs.net Betriebs GmbH, ECLI:EU:C:2017:218, para. 42.

[44] See e.g. Enrico Bonadio and Nicola Lucchi, "How Far Can Copyright Be Stretched? Framing the Debate on Whether New and Different Forms of Creativity Can Be Protected, Intellectual Property Quarterly 115 (2019) (discussing the applicability of copyright to AI-generated works and the pressure such technologies place on traditional concepts of authorship and originality).

[45] See Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on Copyright and Related Rights in the Digital Single Market and Amending Directives 96/9/EC and 2001/29/EC, Official Journal of the European Communities 2019 L 130, 92

[46] See e.g. Marcus von Welser, Generative KI und Urheberrechtsschranken, GRUR-Prax 516 2023 (arguing that the opt-out system is poorly designed and ineffective, especially given the lack of reliable enforcement tools like robots.txt); Giuseppe Abbamonte, The Application of the Copyright TDM Exceptions and Transparency Requirements in the AI Act to the Training of Generative AI, 46 E.I.P.R. 479 (2024) (highlighting the legal and technical complexity of implementing machine-readable opt-outs and the risk of undermining Article 4); Tim W. Dornis, The Training of Generative AI is Not Text and Data Mining, 47 E.I.P.R. 65 (2025) (criticizing the extension of the TDM exception to generative AI and warning that the opt-out fails to address the underlying incompatibility); Matthias Leistner, TDM und KI-Training in der Europäischen Union, GRUR 1665 (2024) (noting that the opt-out mechanism is likely to become the key challenge in EU copyright law and is currently unworkable in practice); Gina Maria Ziaja, The Text and Data Mining Opt-Out in Article 4(3) CDSMD: Adequate Veto Right for Rightholders or a Suffocating Blanket for European Artificial Intelligence Innovations?, 19 J. Intell. Prop. L. & Prac. 453 (2024) (arguing that the opt-out introduces uncertainty and may hinder AI development in the EU); Adam Buick, Copyright and AI Training Data—Transparency to the Rescue?, 20 J. Intell. Prop. L. & Prac. 182 (2025) (explaining that transparency obligations under the AI Act cannot resolve the structural flaws of the opt-out mechanism under Article 4); Rossana Ducato & Alain Strowel, Ensuring Text and Data Mining: Remaining Issues With the EU Copyright Exceptions and Possible Ways Out, CRIDES Working Paper No. 1/2021, at 4–7 (noting that legal uncertainty, technical blocks, and the complexity of opt-out implementation may frustrate the legislative intent behind Article 4).

generate further legal uncertainty for developers and users of AI-generated content, as they struggle to determine whether their training activities comply with copyright law.[47]

In the opinion of the author, it is a stretch to say that the two TDM exceptions in the CDSM Directive are fit for purpose when applied to the development of AI systems. The original intent of Articles 3 and 4 was to promote research and innovation, but they were not specifically designed to address the scale, complexity, or technological architecture of modern AI training pipelines.[48] As such, we will highlight several reasons (see Section 2.1 of this study) why the current legal framework may fall short in addressing the needs of generative AI. These include the mismatch between the broad use of copyright-protected works in AI training and the restrictive scope of permitted uses, the ambiguity surrounding opt-out declarations, and the lack of legal certainty around the status and use of AI-generated outputs.

Another important feature of EU copyright law is that it remains rooted in a human-centric vision of creativity. Unlike the UK, which recognises "computer-generated works" under Section 9(3) of its Copyright, Designs and Patents Act 1988 and assigns authorship to the person making the necessary arrangements,[49] the EU approach insists on a direct link to human creativity.[50] However, determining

---

47 See e.g Severine Dusollier et al., Copyright and Generative AI: Opinion, 16 JIPITEC 121 (2025) (arguing that the lack of clarity concerning the technologies, modalities, timing, and location for expressing the opt-out under Article 4(3) CDSM contributes to legal uncertainty and should be urgently addressed).

48 Ibidem (noting that Articles 3 and 4 CDSM were enacted before the emergence of generative AI and may not cover all aspects of AI model development and operation). See also European Commission and Jean-Paul Triaille et al., Study on the Legal Framework of Text and Data Mining (March 2014), available at https://data.europa.eu/doi/10.2780/1475 (clarifying that the legal concept of TDM was originally tailored to support research-oriented data analysis); See Commission Staff Working Document, Impact Assessment on the Modernisation of EU Copyright Rules Accompanying the Document Proposal for a Directive of the European Parliament and of the Council on Copyright in the Digital Single Market and Proposal for a Regulation of the European Parliament and of the Council Laying Down Rules on the Exercise of Copyright and Related Rights Applicable to Certain Online Transmissions of Broadcasting Organisations and Retransmissions of Television and Radio Programmes, SWD(2016) 301 final (Brussels, 14 September 2016) (the document clearly shows that the TDM exceptions were conceived as an experimental policy mechanism aimed at balancing data-driven innovation and IP protection in narrowly defined contexts—such as bioinformatics, medical research, and textual analysis for knowledge discovery—with the primary objective of fostering European research competitiveness, not enabling large-scale AI training. Furthermore, the documents consistently stress the narrow interpretation of exceptions under the three-step test in international and EU copyright law, which would exclude expansive uses such as model training unless explicitly authorised).

49 See § 9(3) of the UK Copyright, Designs and Patents Act 1988. For a more detailed discussion, see E Bonadio et al, Will Technology-Aided Creativity Force Us to Rethink Copyright's Fundamentals? Highlights from the Platform Economy and Artificial Intelligence (2022) 53 International Review of Intellectual Property and Competition Law 1174, 1187. But see contra Matt Blaszczyk, Impossibility of Emergent Works' Protection in U.S. and EU Copyright Law, 25 North Carolina Journal of Law & Technology 1, 15–20 (2023) (arguing that this provision conflicts with Section 1 of the same Act, which limits copyright protection to "original literary, dramatic, musical, or artistic works." Blaszczyk observes that while the UK Copyright Act attributes authorship of computer-generated works to the person making the necessary arrangements, Section 1's originality requirement creates a fundamental tension. This inconsistency, he contends, mirrors the inherent conceptual paradox of "emergent" or "authorless" works: absent human authorship, there can be no original expression of ideas, and thus no copyrightable subject matter. In his view, the statutory framework for computer-generated works is logically irreconcilable with copyright law's doctrinal foundations).

50 In addition to the constant case law of the CJEU affirming the human-centric concept of authorship, see also European Parliament Resolution of 20 October 2020 on intellectual property rights for the development of artificial intelligence

who qualifies as the "arranger" is not always straightforward and typically requires a case-by-case assessment. The UK provision reflects a much earlier technological context—one in which "computer-generated" referred to deterministic outputs from narrowly programmed systems.[51] By contrast, contemporary AI models exhibit far greater complexity and autonomy, making it increasingly unclear who, if anyone, is meaningfully responsible for the "arrangements" behind a given output. As a result, the UK's approach, though formally clear, may be poorly suited to address the realities of modern generative systems and offers limited guidance for evaluating authorship in today's hybrid human–machine creative processes.[52] The United States follows a similar human authorship requirement and has explicitly excluded AI-generated works from copyright protection unless there is meaningful human input involved.[53] Some EU stakeholders and academics are exploring alternative frameworks – such as sui generis or neighbouring rights – for certain types of AI-generated content to balance innovation incentives with legal coherence.[54] While no consensus has yet emerged, these discussions indicate a willingness to consider tailored solutions beyond the traditional copyright paradigm.

In addition to originality and authorship, EU copyright is also guided by the principle of proportionality.[55] This principle seeks to balance the rights of creators with the broader needs of society. Exceptions to copyright—for education, private use, public interest reporting, and now data mining—are meant to ensure that copyright does not become a barrier to access, research, and innovation. However, when it comes to generative AI, this balance is increasingly difficult to achieve.

Finally, EU copyright law operates within a broader international context, shaped by agreements such as the Berne Convention[56] and the TRIPS Agreement.[57] These establish baseline protections that all

technologies (2020/2015(INI)), 2021 O.J. (C 404) 129, at §8 (affirming that copyright protection should only be granted to intellectual creations that are human-made and that the concept of authorship is inherently linked to natural persons).

51 See Intellectual Property Office (UK), "Artificial intelligence call for views: copyright and related rights" (UK Government, 2020) https://www.gov.uk/government/consultations/artificial-intelligence-and-intellectual-property-call-for-views

52 For some additional critical comments on this provision, see P. Bernt Hugenholtz, and Joao Pedro Quintais, Copyright and Artificial Creation: Does EU Copyright Law Protect AI-Assisted Output? (2021) 52 International Review of Intellectual Property and Competition Law 1190, 1211 (noting that "since the introduction of the regime on computer-generated works in UK law in 1988, this has led to just a single court decision, which has not clarified this issue").

53 According to the current version of the *Compendium of U.S. Copyright Office Practices*, copyright protection will be refused if a human being did not create the work—such as when a machine operates autonomously or randomly, without meaningful human input or intervention, or when the work is created by a non-human animal. See *U.S. Copyright Office, Compendium of U.S. Copyright Office Practices* §§ 101, 306, 312.2 (3d ed. 2021); see also *Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence*, 88 Fed. Reg. 16190 (Mar. 16, 2023).

54 See Council of the European Union, "Policy Questionnaire on the Relationship Between Generative Artificial Intelligence and Copyright and Related Rights," Document ST 16710/1/24 REV 1, 2024, pp. 18–19; Ana Ramalho, Will Robots Rule the (Artistic) World? A Proposed Model for the Legal Status of Creations by Artificial Intelligence Systems, 21 Journal of Internet Law 1 (2017) (proposing a sui generis neighbouring right model without human authorship).

55 See, e.g., Case C-201/13, Deckmyn v Vandersteen, ECLI:EU:C:2014:2132, para. 27; Case C-314/12, UPC Telekabel Wien, ECLI:EU:C:2014:192, paras. 46–63; Case C-70/10, Scarlet Extended, ECLI:EU:C:2011:771; and Case C-360/10, Netlog, ECLI:EU:C:2012:85.

56 Berne Convention (Berne Convention for the Protection of Literary and Artistic Works, 9 September 1886, S. Treaty Doc. 99-27, 1161 U.N.T.S. 3 (amended 2 September 1979).

57 Agreement on Trade-Related Aspects of Intellectual Property Rights, 15 April 1994, Marrakesh Agreement Establishing the World Trade Organization, Annex 1C, 1869 U.N.T.S. 299 (TRIPS).

signatories must respect but leave room for national and regional variations. Within this international framework, the EU has traditionally emphasised strong author rights and cultural diversity. As AI technologies evolve, these values will need to be reassessed and possibly reinterpreted to meet new challenges.

While the EU copyright framework provides a strong foundation for protecting human creativity, certain uses of generative AI expose legal uncertainties that may require targeted enforcement and, where necessary, principled reform. The principles of originality, human authorship, and proportionality remain central—but they are being tested by new modes of content creation that blur the lines between input, output, and authorship. The TDM exceptions introduced by the CDSM Directive represent an important first step, but their coverage of generative AI training remains highly contested due to differences in purpose, scale, and legal interpretation.[58] In light of these challenges, the EU may need to consider more targeted legislative or enforcement responses to ensure that its copyright system remains both relevant and robust also in the age of artificial intelligence.

1.4. The challenge: copyright law and generative AI

The rapid development and deployment of generative AI systems pose fundamental challenges to the existing copyright framework in the European Union. As we have seen, these systems, which include large language models (LLMs), image generators, and music composition tools, rely on vast datasets—often scraped from online sources—that include a wide range of protected works. This "training" phase, essential to the AI's performance, typically involves reproducing, storing, and analysing millions of works, many of which are covered by copyright. However, current EU copyright law was not designed with this scale or technological architecture in mind, and as a result, key aspects of the legal framework are under significant strain.

One of the main points of tension is the use of the text and data mining (TDM) exceptions in the Copyright in the Digital Single Market Directive. While Article 3 allows TDM for scientific research by non-commercial entities, Article 4 was meant to enable broader access, provided that rightsholders do not opt out. But this opt-out mechanism, as many scholars have pointed out, may undermine the exception's practical utility in the AI context.[59] If widely applied, the opt-out can render datasets

[58] See e.g. District Court of Hamburg, Robert Kneschke v. LAION e.V., Case No. 310 O 227/23; GEMA v OpenAI, LLC and OpenAI Ireland Ltd. Available at https://www.gema.de/en/w/gema-files-lawsuit-against-openai; Gema v Suno Inc., Available at https://www.gema.de/en/w/press-release-lawsuit-against-suno; SNE, SGDL and SNAC v Meta, Available at https://www.sne.fr/press-release-authors-and-publishers-unite-in-lawsuit-against-meta-to-protect-copyright-from-infringement-by-generative-ai-developers; DPG Media et al. v. HowardsHome, Rechtbank Amsterdam, C/13/737170 / HA ZA 23-690, ECLI:NL:RBAMS:2024:6563 (October 30, 2024); Municipal Court of Appeals of Budapest, Case 9.Pf.20.353/2024/6-II, 3 December 2024. See also the recent first referral at the CJEU, See CJEU, Case C-250/25, Like Company v. Google Ireland, preliminary reference lodged on 3 April 2025. Referral from Fővárosi Törvényszék (Budapest Metropolitan Court), Hungary. Available at https://curia.europa.eu/juris/liste.jsf?num=C-250/25&language=en.

[59] See e.g. João Pedro Quintais, Generative AI, copyright and the AI Act, 56 Computer Law & Security Review 1-17 (2025) (warning that the Article 4 opt-out may undermine the practical utility of the TDM exception in AI training contexts); Thomas Margoni & Martin Kretschmer, A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology, 71 GRUR Int'l 685, 687–89 (2022) (arguing that the opt-out mechanism under Article 4 CDSM undermines the effectiveness of the exception); Christophe Geiger, Giancarlo Frosio & Oleksandr Bulayenko, Text and Data Mining: Articles 3 and 4 of the Directive 2019/790/EU, in Propiedad Intelectual y Mercado Único

incomplete or legally risky, impeding innovation while failing to provide clear protection for rightsholders. Furthermore, the technical and legal uncertainty around how rights are to be reserved in a machine-readable way has led to inconsistent implementation, making it difficult for developers to know whether their use of data is lawful.

In addition, the current rules provide little guidance on the status of AI-generated outputs. EU copyright law is premised on the idea of human authorship, meaning that outputs automatically processed by AI—no matter how complex or human-like they may appear—do not qualify for protection unless a human has made significant creative contributions. This leads to a grey zone in which high-value content may fall outside the scope of protection altogether, raising concerns about ownership, liability, and reuse. Moreover, the legal separation between the input (used for training) and the output (synthetically produced content) does not reflect the reality of AI development, where the two are deeply interlinked. If outputs are substantially based on or resemble training data, questions of copyright infringement may arise, but there is little guidance in current legislation or case law on how to assess this.

In light of these limitations, this study will explore a series of fundamental questions that aim to inform future EU policy. How can copyright law strike a fair balance between protecting creators and enabling innovation in the age of AI? Do the TDM exceptions under Articles 3 and 4 adequately address the scale and nature of AI training, or do they require clarification or revision? Should there be clearer rules for reserving rights and expressing opt-outs in machine-readable formats? How can the EU ensure fair remuneration for righthsolders whose works are used in the development of generative AI? And finally, should the EU consider new categories or mechanisms of protection for AI-assisted or AI-generated works?

These questions reflect the need for a careful, evidence-based reassessment of how copyright law can remain relevant, effective, and fair in an era where creative production is increasingly shaped by non-human actors. The answers will have major implications not only for legal certainty and economic development but also for the future of cultural and scientific creativity in Europe.

The relationship between generative AI and copyright law is not one-directional. While AI challenges the adequacy of existing legal frameworks, copyright law itself may constrain the future trajectory of AI technologies.[60] Ongoing litigation—particularly in the United States—and potential shifts in legislative interpretation risk significantly narrowing the permissibility of AI training practices,[61] a trend now also emerging in the EU following the first referral of a generative AI copyright case to the Court

Digital Europeo 27 (Concepción Saiz García & Raquel Evangelio Llorca eds., Tirant lo Blanch 2019) (noting that the opt-out under Article 4 risks undermining the practical effect of the TDM exception).

[60] See Daryl Lim, Generative AI and copyright: principles, priorities and practicalities, 18 Journal of Intellectual Property Law & Practice 841 (2023) (arguing that generative artificial intelligence serves as a stress test for copyright law)

[61] See Pamela Samuelson, Generative AI meets copyright: Ongoing lawsuits could affect everyone who uses generative AI, 381 Science 158-161 (2023). An updated list of lawsuits against generative AI developers can be found in the Database of AI Litigation (DAIL) available at https://blogs.gwu.edu/law-eti/ai-litigation-database/

of Justice.[62] Should plaintiffs succeed, only those generative AI systems trained on public domain works or under licensed conditions might remain lawful, with far-reaching consequences not only for developers but also for a wide range of sectors increasingly reliant on AI innovation.[63] In this context, it is essential to adopt a forward-looking and adaptable policy framework that anticipates the potential systemic ripple effects that such rulings could generate across the entire AI ecosystem—especially if restrictive interpretations begin to exert transnational influence. These dynamics underscore the urgency of a comprehensive and coherent reassessment of EU copyright law to ensure legal certainty and innovation readiness.

[62] See Case C-250/25, *Like Company v. Google Ireland*, pending before the CJEU, which raises questions regarding the reproduction and communication to the public of press content by generative AI systems under Directives 2001/29 and 2019/790.

[63] See Pamela Samuelson, Generative AI meets copyright: Ongoing lawsuits could affect everyone who uses generative AI, cit.

2. USING COPYRIGHT-PROTECTED WORKS TO TRAIN GENERATIVE AI (INPUT SIDE)

KEY FINDINGS:

AI training entails systematic reproduction of protected works: To train generative AI models, developers copy and store vast datasets—including books, music, and images—raising clear copyright concerns under EU and international law.

CDSM Directive's TDM exceptions are misaligned with generative AI: Article 3 applies only to scientific research conducted by eligible institutions—acting on a not-for-profit basis or under a public-interest mission—and cannot be opted out of. Under certain conditions, this may include public-private partnerships. Article 4 permits broader use but allows rightsholders to opt out. This dual regime is ill-suited to large-scale AI training.

Legal ambiguity hinders both innovation and protection: Key terms like "lawful access" and "appropriate opt-out" lack harmonised definitions or technical standards, creating compliance risks for AI developers and enforcement challenges for rightsholders.

Generative AI goes beyond traditional TDM: Unlike standard text and data mining, which focuses on extracting factual patterns or insights, generative AI systems internalise and replicate expressive content. This qualitative difference arguably places generative training outside the intended analytical scope of the current TDM exceptions.

Transparency Measures Alone Cannot Guarantee Compliance: The AI Act requires disclosure of training data summaries but lacks mechanisms for traceability, auditability, or individual rights enforcement.

Scholarly and legal opinion is shifting: An emerging consensus holds that training generative models constitutes reproduction—not mining—making Article 4 an inadequate legal basis for such training activities.

Fragmentation persists across Member States: National implementations differ widely in scope, technical criteria, and enforcement, undermining the goal of a harmonised Digital Single Market.

Rightsholders receive no compensation: Despite the commercial value generated by AI models, there is no remuneration mechanism for authors whose works are used in training—deepening the "value gap."

Other jurisdictions offer alternative models: Japan's non-enjoyment principle, the U.S. fair use framework, and the UK's contractual override ban all offer lessons that could inform EU reform.

New legislative solutions are required: Options include statutory licensing, collective management, and remuneration rights—but these must be carefully designed to ensure fairness, feasibility, and innovation.

As discussed, the impressive capabilities of generative artificial intelligence (AI) systems—whether producing text, images, music, or code—are made possible through a process known as training. This

process requires large volumes of data to be ingested and analysed by machine learning models to identify underlying patterns and relationships. In the case of generative AI, such as large language models or image generators, training typically involves the large-scale copying and storage of diverse content—including books, newspaper articles, songs, photographs and websites—into digital corpora used to build and refine the AI's capabilities. As outlined in Chapter 1, this new section carries out the study's first analytical step: examining whether current generative AI practices—particularly during the training phase—comply with existing EU copyright provisions.

From a copyright perspective, this phase is particularly sensitive. The creation of a training corpus through web scraping or database extraction often entails the prior reproduction and storage of protected works, regardless of whether those works are later recognisable in the model's outputs.[64] Under EU copyright law, this can constitute an act of reproduction within the meaning of Article 2 of the InfoSoc Directive, which confers exclusive rights on authors to authorise or prohibit such copying. A similar rule exists in the United States, where 17 U.S.C. § 106(1) grants rightsholders the exclusive right to reproduce their works in copies or phonorecords.[65]

The key legal question, then, is not whether reproduction has occurred—it has—but whether that reproduction is permissible under a copyright exception or limitation. Both text and data mining (TDM) and generative AI entail acts of reproduction, but under EU law, their permissibility depends on purpose. TDM refers to automated analytical techniques used to extract patterns, trends, or correlations from large datasets, typically for scientific or informational purposes.[66] In this context, the EU introduced two targeted exceptions in the Directive on Copyright in the Digital Single Market (Directive (EU) 2019/790):[67] As previously illustrated, Article 3, which applies to non-commercial research by public-interest institutions, and Article 4, which allows broader uses—including by commercial entities—provided rightsholders have not opted out using machine-readable means. Recital 11 clarifies that Article 3 may extend to public–private partnerships, as long as private partners do not enjoy preferential access to the results. [68]

---

[64] See also on this, U.S. Copyright Office, Copyright and Artificial Intelligence, Part 3: Generative AI Training (Pre-Publication Version, May 2025) at 28. Available at https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf (noting that the creation of a training dataset from copyrighted works "clearly implicate[s] the right of reproduction," thus rendering such acts presumptively infringing in the absence of a valid exception or defence, such as fair use).

[65] 17 U.S.C. § 106(1) provides that "the owner of copyright under this title has the exclusive rights to do and to authorize [...] to reproduce the copyrighted work in copies or phonorecords." See U.S. Copyright Act of 1976, Pub. L. No. 94-553, § 106, 90 Stat. 2541 (codified as amended at 17 U.S.C. § 106).

[66] For this definition, see Article 2(2) of Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market.

[67] See Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on Copyright and Related Rights in the Digital Single Market and Amending Directives 96/9/EC and 2001/29/EC, Official Journal of the European Communities 2019 L 130, 92.

[68] See Recital 11 of Directive (EU) 2019/790, which provides that "research organisations and cultural heritage institutions should also benefit from such an exception when their research activities are carried out in the framework of public-private partnerships." However, Article 2(2) clarifies that the results of such research cannot be enjoyed on a preferential basis

However, generative AI operates in a very different manner. While TDM systems extract factual or semantic insights from data, generative AI models—such as those based on transformer or diffusion architectures—are trained to synthesise new outputs by encoding and internalising expressive features of the input content.[69] During training, these models build multi-dimensional parameter spaces that allow them to reproduce style, structure, and composition, enabling outputs that closely resemble original creative works.[70] This shift from extraction to expressive recombination marks a significant departure from the analytical logic that underpins the TDM exceptions. This process goes beyond the analytical purpose envisaged by the TDM exceptions, moving instead toward expressive reproduction.[71] In legal terms, it challenges the applicability of Articles 3 and 4 to generative AI training and calls for precise clarification or reform. The TDM exceptions were not designed to accommodate machine-based replication of creative forms on this scale or with this degree of fidelity—nor to serve commercial uses detached from scientific inquiry.[72]

Beyond the question of legal applicability, the use of TDM exceptions for AI training has also become increasingly controversial from a policy and structural standpoint. Many stakeholders argue that the current framework is ill-suited to the scale and nature of generative AI development.[73] Critics highlight

by an undertaking that exercises decisive influence over the organisation—thereby excluding certain forms of commercially-driven research from benefiting under Article 3.

[69] See e.g. Kai Riemer and Sandra Peter, Conceptualizing Generative AI as Style Engines: Application Archetypes and Implications, 79 International Journal of Information Management 1–15, at 2 (2024) (noting that generative AI systems encode essential features or patterns of input data (training data) into what is known as the latent space); See also OECD, Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449 (revised 2024). Available at https://oecd.ai/assets/files/OECD-LEGAL-0449-en.pdf. Section I (providing a general definition of AI systems as machine-based systems that infer from input how to generate outputs such as content or decisions); Andrea Asperti & Valerio Tonelli, Comparing the latent space of generative models, 35 Neural Computing and Applications 3155–3172, at 3164 (2023) (explaining that generative models learn an internal latent space that captures the relevant features of the data, which enables the synthesis of realistic new samples).

[70] See generally Andrea Asperti & Valerio Tonelli, Comparing the latent space of generative models, 35 Neural Computing and Applications 3155–3172(2023) (explaining that during training, generative models construct internal representations—known as latent spaces—that encode and retain key features of the input content. These encoded features are then used to generate new outputs that often replicate the style, structure, or composition of the original data).

[71] By expressive reproduction, this study refers to the internalisation and probabilistic recombination of expressive elements—such as structure, style, or composition—derived from protected works, in a manner that emulates the original's perceptible form.

[72] See footnote 77 & 78 below for further discussion.

[73] See e.g. Joint Letter to Members of the European Parliament on the Impact of Artificial Intelligence on the European Creative Community (23 July 2024), available at https://composeralliance.org/media/1651-joint-letter-to-members-of-the-european-parliament-on-the-impact-of-artific.pdf signed by major European creators' associations, calling for an end to the unlicensed use of protected works in AI training, greater enforcement of authorial consent, and the reform of Article 4 of the CDSM Directive to safeguard creator rights; Creators for Europe United, Open Letter to the European Commission for Fair, Transparent, and Legally Compliant AI Development (25 April 2025), available at https://creators-for-europe-united.eu (highlighting creators' demands for consent, transparency, and fair remuneration in AI training); Open Letter to the Attention of Ministers of Culture Ahead of the Education, Youth, Culture and Sport Council on 12–13 May 2025 (6 May 2025), available at: https://composeralliance.org/media/1864-open-letter-to-the-attention-of-ministers-of-culture-ahead-of-the-education.pdf (endorsed by a broad coalition of organisations representing writers, translators, journalists, performers, composers, visual artists, and screen directors, calling for strong safeguards for

issues such as the ease with which rightsholders can opt out, the lack of harmonised technical standards for implementing such reservations, and the absence of any corresponding remuneration mechanism for content used under Article 4. As AI developers increasingly rely on online content that is not freely reusable, these shortcomings raise broader concerns about fairness, legal certainty, and the sustainability of creative ecosystems.

This chapter examines the legal and policy complexities surrounding the use of copyright-protected content in AI training. It begins by analysing the legal basis for text and data mining (TDM) under the CDSM Directive, clarifying the technical distinction between TDM and generative AI (Section 2.1.1), and assessing whether generative AI training can qualify as TDM under current EU law (Section 2.1.2). It then considers the legal consequences of unauthorised training (Section 2.1.3), the structural gaps in the current framework (Section 2.1.4), and anticipated developments in case law of the CJEU (Section 2.1.5). The chapter concludes this legal overview with a comparative analysis of international TDM regimes and their relevance for EU reform (Section 2.1.6).

The chapter then moves to examine divergent national implementations of the TDM exceptions, which risk creating legal fragmentation within the internal market (Section 2.2). It further explores the concerns of rightsholders—particularly in relation to control over their works and the lack of compensation mechanisms (Section 2.3)—and discusses ongoing debates around author's rights and remuneration models for AI training (Section 2.4). Finally, it analyses the interface between copyright and the newly adopted Artificial Intelligence Act, with particular attention to transparency obligations for developers of general-purpose AI systems (Section 2.5).

By linking copyright exceptions, implementation inconsistencies, rightsholders concerns, and regulatory responses, this chapter offers a structured overview of one of the most pressing and controversial aspects of the intersection between generative AI and copyright law. While the EU has taken significant steps to update its legal framework, major gaps persist—not only in terms of legal clarity, but also with respect to the equitable allocation of value in a rapidly evolving, AI-driven creative economy.

2.1. Text and Data Mining (TDM) in the CDSM Directive

TDM is defined in Article 2(2) of the CDSM Directive as "any automated analytical technique aimed at analysing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations."

Until recently, the debate surrounding the use of copyright-protected content for training generative AI systems in Europe largely revolved around the applicability of the text and data mining (TDM) exceptions introduced by the CDSM Directive. Article 4,[74] in particular, has been at the center of

copyright and transparency under the AI Act and condemning the unauthorised use of members' works and data for AI training without consent or remuneration).

[74] Article 4 was introduced during the legislative process through Amendment 65 of the JURI Committee Report (Rapporteur: Axel Voss), which proposed an optional TDM exception for users with lawful access, provided that rightholders had not expressly reserved their rights, including via machine-readable means. See: European Parliament, Committee on Legal Affairs, *Report on the Proposal for a Directive of the European Parliament and of the*

attention, seen by many as the legal gateway for AI developers—especially commercial actors—to scrape and process massive volumes of online content.[75] However, a shift is occurring in academic debate. A growing number of scholars have begun to argue that training generative AI systems does not qualify as TDM, either technically or legally, or is at least highly problematic.[76] Their concern is that generative AI does not merely extract knowledge from data—it synthesises digitally processed content that may directly compete with the original works, such as images, music, or text. This perspective is gaining traction and deserves careful consideration (see Section 2.1.2). While compelling, the assertion that commercial AI training uniformly falls outside the scope of Article 4 may overstate a legal position that remains unsettled both in case law and national practice. A more balanced interpretation recognises that, without harmonised EU guidance, legal uncertainty prevails. At the same time, the fact that new technologies disrupt traditional business models does not by itself justify broadening copyright protection.[77] Before turning to the specific legal uncertainties surrounding Article 4, it is useful to consider how this provision came to be regarded as a potential legal basis for generative AI training. This interpretation has gained traction primarily in the absence of a dedicated legal framework regulating the ingestion of protected works for AI development. Rather than emerging from established legal doctrine or jurisprudence, the view that Article 4 permits such practices has developed through a combination of textual ambiguity, regulatory silence, and widespread industrial reliance.

First, the apparently broad and technologically neutral definition of text and data mining (TDM) under Article 2(2), combined with the open-ended language of Article 4, has been interpreted as offering implicit coverage for large-scale data uses, even when involving expressive works. **Second**, due to the

Council on Copyright in the Digital Single Market (COM(2016)0593 – C8-0383/2016 – 2016/0280(COD)), Amendment 65.

[75] See e.g. Martin Senftleben, The TDM Opt-Out in the EU – Five Problems, One Solution: Why the EU Should Introduce a Remuneration Right for Text and Data Mining Instead of Relying on the Rights Reservation Option under Article 4 CDSMD, Kluwer Copyright Blog, 21 February 2024 (arguing that Article 4 has become the de facto legal basis for commercial AI training in the EU, despite not having been designed for this purpose), available at: https://copyrightblog.kluweriplaw.com/2025/04/22/the-tdm-opt-out-in-the-eu-five-problems-one-solution/

[76] See e.g. Tim Dornis, The Training of Generative AI Is Not Text and Data Mining, 47 European Intellectual Property Review, 65-78 (2025); Tim Dornis, Generative AI, Reproductions Inside the Model, and the Making Available to the Public. IIC - International Review of Intellectual Property and Competition Law (2025); Schack, Haimo Auslesen von Webseiten zu KI-Trainingszwecken als Urheberrechtsverletzung de lege lata et ferenda 77 NJW - Neue Juristische Wochenschrift 113-118 (2024) at § 8 (2024); Welser, Marcus, Generative KI und Urheberrechtsschranken, GRUR-Prax 516-520 at § 19 (2023); Baumann, Malte, Generative KI und Urheberrecht – Urheber und Anwender im Spannungsfeld, NJW - Neue Juristische Wochenschrift 3673-3678 at § 14 (2023); Jonathan Pukas, KI-Trainingsdaten und erweiterte kollektive Lizenzen: Generierung von Werken als KI-Trainingsdaten auf Basis erweiterter kollektiver Lizenzen, GRUR 2023, 614 (strengthening the argument that the current TDM exceptions are conceptually unsuited for the training of generative AI models); Bob Brauneis, Copyright and the Training of Human Authors and Generative Machines, 48 Columbia Journal of Law and the Arts 1 (2025); Matthew Sag and Peter K. Yu, The Globalization of Copyright Exceptions for AI Training, 74 Emory Law Journal, (2025) (distinguishing non-expressive use (e.g., AI training) from classical TDM and noting that generative AI reproduces vast volumes of copyrighted material in ways that exceed traditional mining (e.g., extracting facts); Eleonora Rosati, Copyright as an Obstacle or an Enabler? A European Perspective on Text and Data Mining and its Role in the Development of AI Creativity, 27 Asia Pacific Law Review 198–217 (2019) (arguing that the use of TDM for AI training in the EU remains legally uncertain and heavily restricted, despite the introduction of Article 4 DSM Directive).

[77] See Malte Stieper and Michael Denga, The international reach of EU copyright through the AI Act, Institut für Wirtschaftsrecht (2024) at 7.

lack of specific guidance at the EU level, many developers have proceeded under the assumption that, in the absence of a valid opt-out, their practices are lawful—despite significant uncertainty around key notions such as "lawful access" or "appropriate opt-out." **Third**, this interpretation has been reinforced by pragmatic reliance: major AI companies have already engaged in extensive ingestion of copyright-protected content, often without enforcement or legal challenge, thereby creating a perception of legitimacy that lacks formal grounding. However, **this reading is not supported by authoritative legal analysis**. The application of Article 4 to generative AI training **remains speculative**, and the underlying doctrinal, economic, and structural concerns challenge its compatibility with the objectives and limits of the TDM exceptions under EU copyright law (see Section 2.1.2). This stands in contrast with certain non-EU jurisdictions—notably Japan, where Article 30-4 of the Copyright Act **expressly** permits data analysis for any purpose, including AI training, and the United Kingdom, where policy proposals have considered expanding the TDM exception to cover commercial uses by default (see Section 2.1.6).

Indeed, the TDM exceptions were conceived with very different practices in mind—namely, automated analytical techniques used to extract information from large volumes of text and data, often in support of scientific research or empirical analysis. This original intent is clearly reflected in a 2014 study commissioned by the European Commission, which defines TDM as a set of methods aimed at discovering knowledge from data—without reference to the reproduction, repurposing, or expressive transformation of protected content for model training—and emphasizes that these copyright exceptions must be interpreted narrowly and in line with the three-step test.[78] A close reading of the Impact Assessment accompanying the Commission's 2016 Proposal further confirms that Article 4 was introduced as a **targeted, experimental policy mechanism** to reduce legal uncertainty and facilitate data-driven innovation, particularly for start-ups, SMEs, and tech companies, while still preserving rightsholders' ability to opt out.[79] Unlike Article 3, which is narrowly limited to scientific research organisations, Article 4 was designed to promote broader—but still bounded—economic uses of TDM in the EU digital economy. The Impact Assessment consistently frames the exception around use cases such as bioinformatics, medical research, and textual analysis for knowledge discovery, and it never refers to machine learning, neural networks, or algorithmic training methods.[80] Moreover, the

[78] See European Commission and Jean-Paul Triaille et al., Study on the Legal Framework of Text and Data Mining (March 2014), Available at https://data.europa.eu/doi/10.2780/1475. The study confirms that the original intent of the TDM exceptions introduced by the CDSM Directive was to facilitate automated analysis for scientific and empirical purposes—not to enable the large-scale ingestion of protected content for AI training. TDM is defined as "The automated processing of digital materials, which may include texts, data, sounds, images or other elements, or a combination of these, in order to uncover new knowledge or insights." (at 17), with illustrative use cases including bioinformatics, research on rare diseases, and textual corpora analysis. The report contains no reference to machine learning, neural networks, or generative models. It then explicitly stresses that copyright exceptions must be interpreted narrowly and in accordance with the three-step test under international and EU law. Taken together, these elements indicate that the legal and policy framework governing TDM was not designed to cover the expressive reproduction or transformation of protected works involved in generative AI training.

[79] See Commission Staff Working Document, Impact Assessment on the Modernisation of EU Copyright Rules Accompanying the Document Proposal for a Directive of the European Parliament and of the Council on Copyright in the Digital Single Market and Proposal for a Regulation of the European Parliament and of the Council Laying Down Rules on the Exercise of Copyright and Related Rights Applicable to Certain Online Transmissions of Broadcasting Organisations and Retransmissions of Television and Radio Programmes, SWD(2016) 301 final (Brussels, 14 September 2016).

[80] Ibidem at 158

underlying documents repeatedly stress that all exceptions must be interpreted restrictively and remain subject to the three-step test under international and EU law.[81] Therefore, the idea that Article 4 was intended—or even foreseen—as a legal foundation for training generative AI systems on copyright-protected material is not supported by the legislative history or policy rationale articulated in the preparatory work. In contrast, the training of generative AI models involves the reproduction and internal transformation of expressive content—often in ways that implicate the core of copyright protection itself. While the full implications of this conceptual shift will be explored in more detail in Sections 2.1.1 and 2.1.2, the legal and policy framework still hinges, for now, on the formal TDM provisions under EU law.

As outlined earlier (see Section 1.3), the CDSM Directive provides two distinct TDM exceptions: a narrow one under Article 3 for scientific research by non-profit institutions (pursuant to a mission of public interest),[82] and a broader but opt-outable one under Article 4, which is often invoked in the context of generative AI. These two provisions differ in scope, purpose, and legal implications—especially regarding lawful access, opt-out conditions, and applicability to commercial training. The table below summarises these key differences before turning to the legal uncertainties that continue to surround Article 4 in practice.

Table 1: Comparison of TDM Exceptions in CDSM Directive

Aspect	Article 3 (Scientific Research TDM)	Article 4 (General TDM)
Who can use it?	Research organisations and cultural heritage institutions	Any user (including commercial entities)
Purpose allowed	Scientific research only	Any purpose (commercial and non-commercial)
Lawful access required?	Yes	Yes
Opt-out available to rightsholders?	No (exception is unconditional)	Yes (opt-out via machine-readable means or terms)
Commercial use allowed?	Only where the research is carried out by eligible institutions for a public-interest mission, even in PPPs (see Recital 11). Private partners cannot enjoy preferential access to results.	Yes
Applies to AI training?	Not suitable for commercial AI developers	Primary provision relied upon for AI training, though contested

The legal interpretation of several key elements in Article 4 remains contested. One central ambiguity lies in the notion of "lawful access." While the general consensus is that if a user can view or access content legally (e.g., through a paid subscription or publicly available site), then they may mine it, edge cases remain unclear. For example, can content accessed through a trial account or scraper bypassing a login screen be considered lawfully accessed? Likewise, how should an opt-out be communicated "in a suitable manner"? The Directive suggests examples but provides no standardised format. In practice, this means that different platforms and publishers have adopted inconsistent methods for reserving

[81] Ibidem at 85; 91; 124.

[82] This includes public-private partnerships (PPPs), provided that the private partners do not enjoy preferential access to the results.

rights, further complicating compliance for AI developers. Some use robots.txt protocols, others metadata, and many rely on clickwrap or browse wrap contracts—none of which are universally recognised.[83]

As highlighted in the European Commission's study on EU copyright and access to data, these interpretive uncertainties—combined with the opt-out clause—may significantly limit the practical utility of Article 4 and risk frustrating the exception's intended goal of enabling broad, lawful text and data mining for research and innovation purposes.[84] Without standardised, enforceable norms for rights reservation, and given the growing technical complexity of AI training, the current legal framework risks failing both sides: developers lack clarity, and rightsholders lack effective control.[85]

Moreover, as many scholars have started to underline, it is a stretch to claim that Articles 3 and 4 of the CDSM Directive are "fit for purpose" when applied to generative AI.[86] The opt-out mechanism risks undermining the comprehensiveness of training datasets, while offering limited transparency to rightsholders about how their content is used. These challenges are compounded by the Directive's focus on "extracting information," whereas generative AI arguably transforms and internalises works at a far deeper level—raising doubts about the very applicability of the TDM framework (see Section 2.1.1 and 2.1.2).

These concerns have not gone unnoticed at the EU level. The newly adopted Artificial Intelligence Act introduces additional transparency obligations, including a requirement for providers of general-purpose AI models to disclose summaries of the training data used.[87] While this is a step forward, the

[83] See e.g. Hanjo Hamann, Artificial Intelligence and the Law of Machine-Readability: A Review of Human-to-Machine Communication Protocols and their (In)Compatibility with Article 4(3) of the Copyright DSM Directive, 15 JIPITEC 102-121 (2024).

[84] See Senftleben, Martin Study on EU copyright and related rights and access to and reuse of data (Publications Office of the European Union, 2022, at 86-87 available at https://data.europa.eu/doi/10.2777/78973.

[85] Ibidem at 40 (noting that "If an automated, machine-based processing of relevant terms and conditions is not possible, the rights reservation option is likely to render Article 4 DSMD de facto mute... The burden of rights clearance can easily put an end to the research project as a whole.")

[86] See e.g. Thomas Margoni & Martin Kretschmer, A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology, 71 GRUR Int'l 685 (2022); Martin Senftleben, Compliance of National TDM Rules with International Copyright Law: An Overrated Nonissue? 53 IIC 53, 1477–1505 (2022); Dornis, Tim W. and Stober, Sebastian, Urheberrecht und Training generativer KI-Modelle - Technologische und juristische Grundlagen) (2024); Conseil supérieur de la propriété littéraire et artistique, Rapport de mission relative à la mise en œuvre du règlement européen sur l'intelligence artificielle, 11 décembre 2024 (2024), available at https://www.culture.gouv.fr/fr/Media/medias-creation-rapide/cspla_rapport_ia_template_dec_.2024.pdf; See Juan-Carlos Fernández-Molina and Fernando Esteban de la Rosa, Copyright and Text and Data Mining: Is the Current Legislation Sufficient and Adequate? 24 portal: Libraries and the Academy 653-672 (2024); Hanjo Hamann, Artificial Intelligence and the Law of Machine-Readability: A Review of Human-to-Machine Communication Protocols and their (In)Compatibility with Article 4(3) of the Copyright DSM Directive, 15 JIPITEC 102-121 (2024); Eleonora Rosati, Is text and data mining synonymous with AI training? 19 Journal of Intellectual Property Law & Practice, 851 (2024); Bob Brauneis, Copyright and the Training of Human Authors and Generative Machines, 48 Columbia Journal of Law and the Arts 1 (2025).

[87] See art. 53(1)(c) AI Act (requiring providers of general-purpose AI models to publish a summary of the training data used).

disclosure alone does not seem to be sufficient to address the underlying issues of legal uncertainty, lawful use, and remuneration.[88]

These institutional concerns are echoed in a recent policy document published by the Council of the European Union.[89] While most Member States considered the current EU legal framework generally sufficient to address the challenges arising from the interaction between generative AI and copyright, a majority nonetheless identified practical areas where greater clarity and legal certainty would be necessary to ensure better implementation of the existing acquis.[90] In particular, the most frequently raised issue concerned the application of the text and data mining (TDM) exception and its opt-out mechanism, as introduced by the CDSM Directive, to the AI training process.[91] A substantial number of contributions pointed to persisting uncertainties regarding the applicability of the TDM exception to AI training activities, a concern shared by several national authorities and various stakeholders.[92] Some Member States expressed the view that certain uses of protected works for AI training purposes might fall outside the scope of the TDM exception. Diverging views also emerged regarding the potential establishment of an EU-wide database to enhance legal certainty around the functioning of the opt-out system, although alternative practical measures, such as the development of common standards, were also suggested.

While Article 4 of the CDSM Directive is currently invoked as the main legal basis for commercial AI training activities, this legal foundation is increasingly being questioned also between legal scholars.[93] The remainder of this section will explore these doubts in more depth, beginning with a technical distinction between TDM and generative AI and a fundamental critique that generative AI training does not, in fact, constitute TDM as defined under EU law (Sections 2.1.1 and 2.1.2).

2.1.1. Understanding the technical distinction between TDM and Generative AI

Before examining whether generative AI training qualifies as text and data mining (TDM) under EU copyright law, it is important to understand how these technologies actually work. AI training processes

---

88 See e.g. Adam Buick, Copyright and AI Training Data—Transparency to the Rescue?, 20 J. Intell. Prop. L. & Prac. 182, 183 (2025) (emphasizing that transparency requirements, while necessary, are insufficient to resolve the broader challenges posed by generative AI, and that policymakers must engage with the deeper task of balancing the competing interests of all stakeholders).

89 EU Policy Questionnaire on the Relationship Between Generative Artificial Intelligence and Copyright and Related Rights, ST 16710 2024 REV 1 – NOTE, 20 December 2024.

90 Ibidem.

91 Ibidem.

92 See e.g. Copyright Initiative, Authors and Performers Call for Safeguards Around Generative AI (April 20, 2023), https://urheber.info/media/pages/diskurs/call-for-safeguards-around-generative-ai/069a7d264a-1697140342/authors-and-performers-call-for-safeguards-around-generative-ai_20.4.2023.pdf ; European Composer and Songwriter Alliance/European Writers' Council et al, Joint Statement from Authors' and Performers' Organizations on Artificial Intelligence and the AI Act (February 9, 2023), https://composeralliance.org/media/1136-joint-statement-on-ai-and-the-ai-act.pdf; European Guild for Artificial Intelligence Regulation, Manifesto for AI Companies Regulation in Europe, http://www.egair.eu/resources/EGAIR_Manifesto_EN.pdf

93 See supra note 76.

are deeply technical, and without at least a basic grasp of their structure, regulatory solutions risk being built on flawed assumptions—resulting in rules that are either too rigid or too vague to be effective.

This section therefore introduces a few key concepts from the field of artificial intelligence, explained in clear and accessible terms. It aims to clarify how traditional TDM differs from the techniques used in generative AI systems, and why this distinction is crucial for assessing the scope of existing copyright exceptions. By outlining how these technologies operate in practice, we hope to provide a sound foundation for the legal discussion that follows and help bridge the gap between engineering and legal interpretation.

A good starting point is to distinguish between Text and Data Mining (TDM) and Generative AI (GenAI), which, despite some overlap, serve very different purposes.[94]
TDM belongs to the field of Data Science, which is primarily about analysing existing information. It involves using software to process large volumes of text, images, or other data in order to find patterns—for example, tracking how often a certain term appears in scientific articles. The goal is to extract knowledge from what already exists.

By contrast, Generative AI falls within the broader field of Artificial Intelligence, and more specifically, Machine Learning. Rather than merely analysing data, generative AI systems are engineered to process large datasets and algorithmically synthesise outputs—such as textual sequences, visual renderings, or audio patterns—based on statistical correlations.[95] While both TDM and GenAI rely on large-scale data, they use it in fundamentally different ways. A simple way to remember the difference is: TDM finds patterns; GenAI synthesises new expressions.

This difference has significant legal implications. Under EU copyright law, TDM may fall within certain exceptions—particularly when used for research purposes. But GenAI, because it can synthesise outputs that resemble or incorporate protected works, raises more complex and unsettled legal questions.

To illustrate further, TDM is just one step in a broader process known as Knowledge Discovery in Databases (KDD).[96] This involves selecting the data, cleaning and transforming it, mining it for patterns, and interpreting the results. Importantly, the term "data mining" can be misleading—there's no extraction of raw material, but rather the identification of patterns or correlations. Some TDM methods do make use of Machine Learning tools, but only to improve the analysis—not to create new, expressive outputs.

---

[94] See Stuart Russell & Peter Norvig, Artificial Intelligence: A Modern Approach (4th ed. 2021).

[95] See Ian Goodfellow, Yoshua Bengio & Aaron Courville, *Deep Learning* (2017) § 20; A Radford et al, Language Models Are Unsupervised Multitask Learners (2019) OpenAI Blog, Available at https://insightcivic.s3.us-east-1.amazonaws.com/language-models.pdf; Rishi Bommasani et al., On the Opportunities and Risks of Foundation Models, cit

[96] See Usama Fayyad et al., From Data Mining to Knowledge Discovery in Databases, 17 AI Magazine 37–54 (1996) (defining Knowledge Discovery in Databases (KDD) as the non-trivial process of identifying valid, novel, potentially useful, and understandable patterns in data, encompassing not only data mining algorithms but the entire pipeline from data selection and preprocessing to interpretation and evaluation).

Generative AI, by contrast, is used when the structure behind the data is not clearly known. These systems are designed to reconstruct underlying structures by modelling them statistically and computing outputs that conform to those patterns—for instance, assembling an image or paragraph not present in the original dataset. If the resulting output closely resembles the training data, it indicates that the model has captured and replicated its structural features. But unlike TDM, this is not mere analysis—it is synthesis, and that distinction matters legally which will be analysed in detail in the next section (2.1.2).

One frequently used concept in GenAI discussions is that of a "latent space"—a compressed internal representation of complex data.[97] This can be thought of as a simplified map that helps the system process or organise information. However, latent spaces are not exclusive to GenAI. Many systems, such as image-recognition tools (e.g., those using Convolutional Neural Networks, or CNNs), also rely on them—usually just to classify or group data (e.g., distinguishing between images of cats and dogs), not to produce synthetic outputs that resemble human-created material.[98] So, the mere presence of a latent space does not mean a system is generative.

That said, some architectures—such as autoencoders—use latent spaces to compress and then reconstruct data.[99] While this process is often used for efficiency (e.g., to transmit images with minimal loss), it can also be extended creatively to generate new variations, raising the legal question of whether such outputs constitute a reproduction of protected material.

More advanced GenAI systems, such as Generative Adversarial Networks (GANs) and diffusion models, take generation further.[100] GANs, for example, start with random noise and produce images that a second component (the "discriminator") evaluates for realism. This architecture reflects that the model goes beyond mere data analysis; it is optimised to emulate learned patterns and synthesise outputs consistent with those patterns. Similarly, Large Language Models (LLMs)—such as ChatGPT—use "transformer" architectures to model the statistical structure of language and compute text outputs, word by word, based on patterns found in vast datasets.

---

97 See Ian Goodfellow, Yoshua Bengio & Aaron Courville, *Deep Learning* (2017).

98 See e.g. See Benjamin L.W. Sobel, Elements of Style: Copyright, Similarity, and Generative AI, 38 Harv. J.L. & Tech. 49, 62 (2024) (clarifying that a latent space is a multi-dimensional way to represent data similarities—often not visible in the raw data—and emphasizing that using a latent space does not, by itself, make a model generative); A. Feder Cooper and James Grimmelmann, The Files are in the Computer: On Copyright, Memorization, and Generative AI, 98 Chi.-Kent L. Rev. (forthcoming), at 9, 22–23, available at SSRN: https://ssrn.com/abstract=4803118 (explaining that discriminative models use latent spaces to classify data without generating new outputs, and that when a generative model closely reproduces parts of its training data, it amounts to literal copying under copyright law because the model stores that data in its internal parameters).

99 See Dor Bank, Noam Koenigstein & Raja Giryes, Autoencoders, ARXIV (Mar. 13, 2020), https://arxiv.org/abs/2003.05991

100 See Ian Goodfellow et al., Generative Adversarial Nets, ARXIV (June 10, 2014), https://arxiv.org/abs/1406.2661.

From a legal perspective, the specific model type (GAN, transformer, etc.) matters less than how the model is used. If an AI tool is applied to restore[101] or reconstruct[102] a damaged artwork (for example, in a cultural heritage context), the goal is knowledge extraction. But if the same system is used to produce new commercial artworks in the style of a known artist, it may involve appropriation of protected material.[103] In both cases, reproduction occurs—but only in the first case might that reproduction fall within the scope of the TDM exception due to its strictly analytical objective. The law must therefore consider the intent and context of use, rather than relying solely on labels like "AI" or "TDM." A deeper distinction between training-based architectures and Retrieval-Augmented Generation (RAG) systems—whose outputs raise different legal issues—is explored in Section 2.1.2.5.

It is also worth noting that even non-generative tools within traditional TDM frameworks can blur the line between analysis and creation. A well-known example is The Next Rembrandt project, which used data from hundreds of Rembrandt's paintings to algorithmically produce a new image in his style.[104] While artistic styles are not generally protected under copyright law, this project illustrates how difficult it can be to separate knowledge extraction from creative reproduction in legal terms.

In conclusion, while TDM and GenAI may rely on some shared technical methods, their functions and outputs differ fundamentally. TDM is about extracting insights from data; GenAI refers to the algorithmic synthesis of outputs that mimic or reproduce expressive patterns found in pre-existing content. This distinction is not just technical—it is central to evaluating whether existing copyright exceptions, such as those in the CDSM Directive, are applicable to GenAI training.
A simplified visual summary of these differences is provided in the table below.

Table 2: Differences between TDM and GenAI

Dimension	**Text and Data Mining (TDM)**	**Generative AI (GenAI)**
Field	Data Science	Artificial Intelligence / Machine Learning
Purpose	Extract knowledge from existing data	Compute synthetic outputs based on learned patterns
Output	Insights, patterns, correlations	Text, images, music, etc.

[101] See Caroline Goldstein, Rembrandt's Revered 'Night Watch' Was Cut Up to Fit Through a Door. With A.I., You Can See It Whole for the First Time in 300 Years, ARTNET NEWS (June 23, 2021), https://news.artnet.com/art-world/operation-night-watch-1982686.

[102] See Jo Lawson-Tancred, Can A.I. Reconstruct the Lost Murals of Delacroix?, ARTNET NEWS (March 31, 2025), https://news.artnet.com/art-world/digital-delacroix-ai-2625734.

[103] See Nicole Sales Giles & Sebastian Sanchez, Cancel the Christie's AI Art Auction, OPENLETTER (Feb. 8, 2025), https://openletter.earth/cancel-the-christies-ai-art-auction-f5135435?limit=0. (An open letter signed by artists and curators criticising Christie's for promoting AI-generated works without proper credit or consent from human creators.)

[104] The Next Rembrandt: Blurring the Lines Between Art, Technology and Emotion, MICROSOFT (Apr. 13, 2016), https://news.microsoft.com/europe/features/next-rembrandt/. (This project used a machine learning model trained on digitised data from 346 public domain paintings to generate a new artwork in the style of Rembrandt—illustrating how training data, though legally free to use, can lead to highly expressive and stylistically distinctive outputs.)

Use of Data	Analyses pre-existing datasets	Processes training data to replicate expressive structures
Relation to Copyright	Involves acts of reproduction, but limited to analytical use permitted under specific exceptions	Involves reproduction and synthesis of expressive structures, often exceeding the scope of permitted exceptions
Legal Relevance under CDSM	Covered by TDM exceptions under certain conditions	Not clearly covered; legal uncertainty

2.1.2. Does Generative AI Training really qualify as Text and Data Mining?

2.1.2.1. Legal Interpretation and the Limits of Article 4

Having clarified the fundamental technical differences between traditional text and data mining (TDM) and generative AI systems, we can now turn to the central legal question: does the training of generative models fall within the scope of the TDM exceptions set out in the CDSM Directive? While the Directive permits certain automated uses of protected content for analytical purposes, the application of these provisions to generative AI training is a subject of intense legal and policy debate. The following section explores this controversy, examining whether the legal concept of TDM—as defined in EU copyright law—can accommodate the expressive, synthesis-based nature of generative AI.

The plain language of Article 4 of the CDSM Directive appears to authorise broad TDM activities by any user with lawful access, including commercial entities, unless the rightsholder has opted out. Based on this literal reading, some can argue that generative AI training falls within the exception—provided content is accessed lawfully and no opt-out is in place. However, this interpretation oversimplifies both the legal framework and the underlying technological realities.

While Recitals 2, 3 and 5 of the CDSM Directive underscore the objective of fostering innovation and knowledge-based economic growth, this policy goal must be interpreted in line with the Directive's internal safeguards. Article 4 was intended to remove contractual barriers to large-scale analysis—not to permit commercial-scale ingestion of expressive works for synthetic purposes. As Recital 3 makes clear, the innovation goal of the Directive is not pursued in isolation but in tandem with the need for a well-functioning marketplace for copyright—one that ensures the sustainability of creative sectors while promoting access to content and new technologies. A broad reading of Article 4 that permits AI training would distort its intended scope, effectively turning it into a de facto compulsory licence—something the EU legislator deliberately avoided.

Even if formal conditions under Article 4 are met, including lawful access and lack of opt-out, this does not imply that any automated processing of text or data qualifies as TDM. Article 2(2) of the CDSM Directive defines TDM narrowly as "any automated analytical technique aimed at analysing text and data in digital form in order to generate information." Recital 8 reinforces this analytical scope, referring to the extraction of knowledge, patterns, or trends—not the synthesis of expressive works. Generative AI training does not produce knowledge in the analytical sense foreseen by Article 4 CDSM. Rather than extracting information or identifying patterns for research purposes, it operates by internalising and

synthetically reassembling expressive content. This synthetic function lies beyond the intended purpose of text and data mining exceptions and should not be equated with lawful analytical uses, such as computational linguistic research or scientific discovery.

Moreover, under settled case law from the CJEU (e.g., Infopaq,[105] Pelham[106]), exceptions to copyright must be interpreted strictly. Using protected content to train generative systems—whose outputs emulate creative expression—goes beyond this analytical boundary and falls outside the protection of Article 4.

Although the legal and policy debate in the EU has long treated the question of whether generative AI systems can rely on the text and data mining (TDM) exceptions in the CDSM Directive as a central issue, recent academic commentary and case law increasingly contest this reading.[107] Legal scholars and technologists argue that training generative AI models is not a form of TDM, either from a technical or legal perspective. As Tim W. Dornis powerfully argues in his detailed legal and technical analysis, the processes involved in training generative AI systems go well beyond the boundaries of what the EU law classifies as TDM.[108] Other scholars, commenting on a recent Hamburg District Court ruling[109] regarding the national transposition of the TDM exceptions, have stressed that the CDSM Directive's exceptions are limited to acts of extraction and reproduction for analytical purposes and do not extend to the subsequent training of AI models or the public dissemination of the resulting datasets.[110] Equating TDM with AI training obscures the distinct and additional legal stages involved in commercial model development.[111]

2.1.2.2. The Three-Step Test and Incompatibility with Generative AI

Despite appearances, Article 4 was never meant to justify large-scale ingestion for creative synthesis[112] but only the automated extraction of patterns, correlations, or trends from large datasets to produce new knowledge or informational insights—a technique often associated with scientific research or data analytics.[113] Therefore, its purpose lies in supporting automated extraction techniques typically used in

[105] C-05/08, Infopaq International v. Danske Dagblades Forening (2009) ECLI:EU:C:2009:465 (Infopaq).

[106] Pelham GmbH v Ralf Hütter and Florian Schneider-Esleben (C-476/17) EU:C:2019:624.

[107] See supra note 76.

[108] See Tim Dornis, The Training of Generative AI Is Not Text and Data Mining, 47 European Intellectual Property Review, 65-78 (2025); Tim Dornis and Sebastian Stober, Urheberrecht und Training generativer KI-Modelle. Nomos, Baden-Baden (2024).

[109] District Court of Hamburg, Robert Kneschke v. LAION e.V., Case No. 310 O 227/23.

[110] See e.g. Eleonora Rosati, Is text and data mining synonymous with AI training? 19 Journal of Intellectual Property Law & Practice, 851 (2024); Haimo Schack, «Auslesen von Webseiten zu KI-Trainingszwecken als Urheberrechtsverletzung de lege lata et ferenda» (2024) 77 NJW 113, 114 (written prior to the decision and thus not directly reflecting the court's reasoning)).

[111] See Eleonora Rosati, Is text and data mining synonymous with AI training? cit.

[112] See OECD, Intellectual Property Issues in Artificial Intelligence Trained on Scraped Data, OECD Artificial Intelligence Papers, No. 33 (2025), at 11 (observing that many IP laws, including copyright frameworks, were conceived before the emergence of AI-driven data scraping and generative model training, resulting in significant legal uncertainties).

[113] See e.g. Christophe Geiger et al., Text and Data Mining in the Proposed Copyright Reform: Making the EU Ready for an Age of Big Data?, 49 IIC 814, 818–820 (2018) (explaining that the TDM exception under EU copyright law was intended to enable the extraction of factual information and insights, rather than the reproduction or internalization of protected

research or data analysis. The Directive's language reinforces this interpretation: Article 2(2) defines TDM as an "automated analytical technique," while Recital 8 confirms that the goal is to extract knowledge—such as patterns or trends—from large-scale analysis.[114] It is also relevant to underline that both TDM exceptions under Articles 3 and 4 of the CDSM Directive relate specifically to acts of reproduction (for copyright) and extraction (for sui generis database rights), but only where such acts are undertaken **for the purpose of analysis**. The CJEU has consistently affirmed that exceptions to copyright must be interpreted strictly. Therefore, any act of reproduction or extraction aimed at generating expressive outputs—such as those involved in generative AI training—cannot be sheltered under these exceptions. Put simply, the legal entitlement to reproduce or extract under Articles 3 and 4 cannot be decoupled from the narrow analytic purpose they were designed to serve.

In addition, any application of the Article 4 exception must comply with the "three-step test" codified in Article 5(5) of the InfoSoc Directive and reflected in international law.[115] This test, embedded in international law through the Berne Convention and TRIPS Agreement, functions as a doctrinal safeguard that ensures exceptions remain narrowly defined, purpose-bound, and proportionate to authorial interestst. It requires that exceptions (i) apply only to certain special cases, (ii) do not conflict with the normal exploitation of the work, and (iii) do not unreasonably prejudice the legitimate interests of the rightsholder. Applied to generative AI, serious doubts arise on all three counts. **First**, large-scale ingestion of expressive works for AI training is no longer a special case—it is becoming a systematic industry practice.[116] **Second**, the ability of generative models to replicate the style, structure, or substance of protected works directly undermines normal exploitation channels, such as licensing and derivative markets. As clarified by the WTO panel in the landmark dispute on the U.S. 'homestyle exemption' (§110(5)(B) Copyright Act), even exceptions that serve public interests or offer some form of compensation may still violate the three-step test if they significantly displace the licensing market for the original work.[117] The standard is not whether some value is returned to authors, but whether the normal channel of economic exploitation is impaired. This impairment may also arise in more subtle forms—even when the AI-generated content does not replicate original works verbatim. The key consideration is whether the output serves as a functional equivalent, fulfilling the same user demand that would otherwise lead to access via legitimate, licensed channels. Such functional substitution can materially interfere with normal exploitation by displacing attention, traffic, or revenues, particularly

expression); Matthew Sag and Peter K. Yu, The Globalization of Copyright Exceptions for AI Training, 74 Emory Law Journal, (2025) (stating that storing datasets and converting them into tokenized formats constitutes reproduction under U.S. (and arguably EU) copyright law, particularly when those copies are stable and human-readable with machines).

[114] See Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market, recital 8, 2019 O.J. (L 130) 92 ("Text and data mining makes the processing of large amounts of information with a view to gaining new knowledge and discovering new trends possible.").

[115] See art. 5(5) of Directive 2001/29/EC [2001] O.J. L 167/10; art. 13 of the TRIPS Agreement; and art. 9(2) of the Berne Convention. The three-step test has been authoritatively interpreted by two WTO Dispute Settlement Panel Reports: Doc. WT/DS160/R of 15 June 2000 (copyright), and WT/DS114/R of 17 Mar. 2000 (patents).

[116] See e.g. Shayne Longpre et al. A large-scale audit of dataset licensing and attribution, 6 Nature Machine Intelligence 975–987 (2024).

[117] WTO, Report of the Panel, United States – Section 110(5) of the US Copyright Act, WT/DS160/R (15 June 2000), §§ 6.72 and 6.97.

where the AI-generated output delivers paraphrased, summarised, or stylistically similar content. In these cases, the interference does not depend on literal copying but on the economic role played by the AI output as a substitute—a factor that must be weighed carefully when applying the second prong of the three-step test. **Third**, this scale of unremunerated use, often without transparency or consent, unreasonably prejudices authors' legitimate interests. These factors suggest that even where formal compliance with Article 4 is asserted, **the three-step test likely fails**—rendering such uses incompatible with EU copyright law. In practical terms, generative AI training fails each step of the test: it is industrial in scale rather than exceptional; it substitutes rather than complements normal exploitation; and it compromises rightsholders' interests through unlicensed and opaque use. Importantly, the test for whether an exception interferes with 'normal exploitation' must account not only for quantitative substitution but for the normative value of licensing channels that incentivise future creation.[118] AI training that captures expressive features to generate competing outputs undermines both dimensions—resulting in a twofold violation of this prong. As a result, Article 4 cannot be relied upon to justify the ingestion of protected works for generative training—legally, such use falls outside the EU exception framework.

2.1.2.3. Technical Structure and the Internalisation of Expression

These legal concerns are mirrored in the technical structure of generative AI. Generative models are not merely analysing data—they are trained to encode and simulate the expressive dimensions of creative works. As previously discussed, they do not simply extract patterns; they internalise and model stylistic and structural elements in order to generate outputs that may closely resemble original expressions. In other words, these systems go beyond mining—they absorb and reorganise protected content into new, synthetic forms. [119]

This interpretation is further supported by the recent U.S. Copyright Office report on training data, which explicitly rejects both the idea that AI training is non-expressive and the analogy to human learning. [120] The Office emphasises that generative models ingest and reproduce expressive forms—not merely factual information—and process them with a mechanical scale and precision that far exceeds human cognition. [121]

---

[118] Ibidem.

[119] See e.g. Benjamin L. W. Sobel, Artificial Intelligence's Fair Use Crisis, 41 Colum. J.L. & Arts 45 (2017) (arguing that machine learning models can internalize and repurpose expressive features of works, not just extract facts); Weijie Huang & Xi Chen, Does Generative AI Copy? Rethinking the Right to Copy Under Copyright Law, 56 Computer L. & Sec. Rev. 106100 (2025) (explaining that GenAI models encode expressive structures such as grammar, style, and tone, rather than merely extracting factual data).

[120] See U.S. Copyright Office, Copyright and Artificial Intelligence, Part 3: Generative AI Training, cit. T 47–48. The Office rejects both the claim that training is purely statistical and the analogy to human learning. It emphasizes that models learn how words and images are selected and arranged—"the essence of linguistic expression"—and absorb creative patterns specifically to replicate them. It also stresses that generative AI training involves creating perfect copies and analysing them at "superhuman speed and scale," unlike humans, who retain only imperfect impressions. These differences are considered fundamental to the fair use analysis.

[121] Ibidem.

Critically, this transformative process stands in stark contrast to the concept of text and data mining (TDM) under the CDSM Directive. Deep learning models do not simply identify patterns for analytical purposes; they learn hierarchical representations of expressive works—including syntax, style, and compositional structure—which they recombine into autonomous outputs.[122] This expressive recombination exceeds the analytical scope defined under Articles 3 and 4 of the Directive, which were designed to support scientific research and information extraction—not machine-led emulation of human creativity

2.1.2.4. Reproduction Right, Memorisation, and Empirical Evidence

A further misconception lies in the assumption that if AI systems do not "store" works in a human-readable format, the reproduction right is not engaged. The Infopaq ruling confirms that even transient copies—if integral to the process and allowing perception—may qualify as reproduction under Article 2 InfoSoc.[123] SAS Institute added that the form or visibility of the reproduction is immaterial; what matters is whether expression is reproduced.[124] Generative AI models encode expressive works during training, transforming them into vector spaces and model weights. This internalisation allows for later output that mimics protected expression. Empirical studies confirm that models can memorize and reproduce content verbatim.[125] This process constitutes a functional equivalent of partial reproduction, even where the output is not identical. Even compressed and abstracted representations in model weights can amount to reproductions if they enable the reconstitution of protected elements. This reflects the technology-neutral and functional interpretation of 'reproduction' under EU law.

Originality resides in the specific form of expression, not in abstract ideas or data. Accordingly, the creation of training corpora through large-scale scraping or data harvesting implicates the reproduction

---

[122] See e.g. Bengio, Yoshua; Lecun, Yann; Hinton, Geoffrey, Deep learning for AI, 64 Communications of the ACM 58-65 (2021) (noting that deep networks "exploit a particular form of compositionality in which features in one layer are combined in many different ways to create more abstract features in the next layer"); Goldberg, Yoav. A Primer on Neural Network Models for Natural Language Processing. 57 Journal of Artificial Intelligence Research, 345–420 (2016) (noting how deep learning learns abstract, high-dimensional, hierarchical features that reflect the underlying structure of language, and then leverages those features to generate expressive and novel outputs).

[123] C-05/08, Infopaq International v. Danske Dagblades Forening (2009) ECLI:EU:C:2009:465 at §40, 42.

[124] Case C-406/10, SAS Institute Inc. v. World Programming Ltd. (ECLI:EU:C:2012:259) at §33. See also Opinion of Advocate General Bot, delivered on 29 November 2011, Case C-406/10, SAS Institute Inc. v. World Programming Ltd., at § 106-107; 119-120.

[125] See e.g. Nicolas Carlini et al. "Extracting training data from diffusion models". 32nd USENIX Security Symposium 5253-5270 (2023) (showing that diffusion models memorize individual images from their training data and emit them at generation time); Nicolas Carlini et al., Extracting training data from large language models. In 30th USENIX Security Symposium (USENIX Security 21) 2633-2650 (2021) (demonstrating that GPT-2 can be prompted to reproduce verbatim paragraphs from its training data — especially when data is duplicated or rare); Jing Huang, et al., Demystifying verbatim memorization in large language models, Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 10711–10732 (2024) (noting that verbatim memorization is intertwined with the LM's general capabilities); Vitaly Feldman, Does learning require memorization? a short tale about a long tail. In Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, pages 954–959 (2020) (observing that a generalization gap in the long tail implies that rare examples must be memorized in order to learn effectively); Vitaly Feldman and Chiyuan Zhang, What neural networks memorize and why: Discovering the long tail via influence estimation. 33 Advances in Neural Information Processing Systems, 2881–2891, (2020); Tim Dornis and Sebastian Stober, Urheberrecht und Training generativer KI-Modelle. Nomos, Baden-Baden (2024).

right under Article 2 of the InfoSoc Directive, as it involves the unauthorised copying and storage of protected works—irrespective of whether these are directly recognisable in the outputs. This foundational misalignment between AI training practices and the TDM exceptions underscores the need for a clearer legal framework that distinguishes between permissible analytical uses and infringing reproductive processes.

This distinction is not merely semantic; it has direct legal consequences. Copyright protects the expression of ideas—not the ideas themselves—by safeguarding the specific form in which a work is written, composed, or visualised. As noted above, TDM operates within a narrow analytical scope focused on semantic extraction, while generative AI systems are designed to internalise and recombine expressive structures. This structural mismatch lies at the heart of the legal and normative concerns under EU copyright law. Generative AI, however, is syntax-hungry: its performance depends on absorbing and reproducing the very elements that copyright is designed to protect. These concerns are not limited to the training stage. According to the recent Report of the U.S. Copyright Office, there is a strong argument that copying a model's weights may implicate the reproduction right when those weights embed memorised examples of protected content.[126] The implications are considerable: if protectable expression is indeed embedded in a model's parameters, "subsequent copying of the model weights, even by parties not involved in the training process, could also constitute prima facie infringement."[127]

2.1.2.5. RAG Systems, Legal Uncertainty, and the Case for Reform

The growing complexity of AI systems—and the evolving ways in which they process, internalise, and reuse data—has led to increased legal uncertainty around the applicability of the TDM exceptions. This uncertainty is particularly visible in attempts to distinguish between generative model training and alternative technical architectures such as Retrieval-Augmented Generation (RAG).[128] At the same time, scholars and policymakers are beginning to challenge the continued reliance on Article 4 of the CDSM Directive as a legal basis for large-scale ingestion of expressive works.

Dornis explains that this misuse of the TDM label stems from a fundamental misunderstanding.[129] Many assume that generative AI, like traditional TDM, only processes semantic information (e.g., facts, themes, or trends). But AI models do not distinguish between semantics and syntax. Technically, they treat all input—whether factual or expressive—as data to be processed. During training, the entirety of

---

[126] See U.S. Copyright Office, Copyright and Artificial Intelligence, Part 3: Generative AI Training, cit. at 28–29.

[127] Ibidem. (suggesting that downstream actors—such as those who fine-tune, distribute, or deploy a model—may also face liability if the model weights embed protectable expression, thus extending potential infringement beyond the training phase).

[128] Retrieval-Augmented Generation (RAG) enables generative AI models to access external data sources—such as online encyclopedias, websites or databases—at the time of a query, incorporating retrieved information into their responses. This allows them to generate context-relevant outputs without requiring prior training on the referenced material. See Patrick Lewis et al., Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, in Advances in Neural Information Processing Systems 33 (2020), at 9459–60, available at https://arxiv.org/abs/2005.11401.

[129] See Tim Dornis, Generative AI, Reproductions Inside the Model, and the Making Available to the Public. IIC - International Review of Intellectual Property and Competition Law (2025); Tim Dornis, The Training of Generative AI Is Not Text and Data Mining, cit.

the content, including stylistic and structural elements, is encoded in what is called a vector space, a kind of compressed internal representation that allows the model to later compute new outputs that echo the original content. These internal vector mappings do not merely analyse data—they encode it in a way that facilitates synthetic reproduction. In copyright terms, this represents a form of reproduction, not just analysis.

Recent empirical studies further confirm that generative AI systems are capable of memorizing and reproducing parts of their training data verbatim, highlighting that expressive content is not just analysed but internalized in ways that implicate the reproduction right.[130] This internalisation enables models to reproduce styles, tones, and structures that copyright law seeks to protect. This differs fundamentally from traditional TDM activities, which do not require such internalization or expressive replication.[131] As several legal scholars have begun to emphasize, equating the ingestion and internalization of expressive content by generative AI systems with traditional analytical or informational uses reflects a profound misunderstanding of both technological realities and copyright principles.[132] A recalibration of the legal framework—not a reinterpretation of outdated exceptions—is therefore required to properly account for the implications of AI training on protected content. Even by human authors learning from copyrighted works is subject to certain copyright limitations; therefore, extending an expansive exception to generative AI systems that internalize expressive elements would lack a sound legal basis.[133] As such, the notion that training processes fall within the safe harbour of text and data mining exceptions must be definitely reconsidered in light of both technological reality and empirical evidence. This analysis should not, however, be conflated with the dinstinct copyright implications of Retrieval-Augmented Generation (RAG) systems. Unlike model training—which entails the reproduction and internalisation of expressive works to adjust model parameters—RAG systems

[130] See supra note 119.

[131] See e.g. Tim Dornis, Generative AI, Reproductions Inside the Model, and the Making Available to the Public, cit. (critizing the misapplication of TDM exceptions to generative AI, explaining that unlike TDM, GenAI involves ingesting and re-expressing copyrighted material in ways that are closer to reproduction than analysis).

[132] See e.g. Bob Brauneis, Copyright and the Training of Human Authors and Generative Machines, 48 Columbia Journal of Law and the Arts 1 (2025) (arguing that generative AI does not simply "analyze" works -as TDM would allow- but absorbs and internalizes the expressive structure -syntax, style, tone- of works —which leads to outputs based on protected expressive elements); Tim Dornis, The Training of Generative AI Is Not Text and Data Mining, cit; Haimo Schack, Auslesen von Webseiten zu KI-Trainingszwecken als Urheberrechtsverletzung de lege lata et ferenda, cit; Mark A. Lemley & Bryan Casey, Fair Learning, 99 Texas Law Review 743 (2021) (arguing that training AI models on copyrighted works requires a separate copyright analysis, distinct from human learning); Nicola Lucchi, ChatGPT: A Case Study on Copyright Challenges for Generative AI Systems, Eur. J. Risk Regulation 1, 11 (2024) (observing that AI systems cannot learn from art in the same way humans do, since they require an exact copy of the artwork in their training dataset); Kalpana Tyagi, Copyright, text & data mining and the innovation dimension of generative AI, 19 Journal of Intellectual Property & Practice 557, 562-63 (2024) (stressing that GenAI "digests" expression, not just semantics and acknowledging that training large models typically requires reproduction of entire or substantial parts of copyrighted works, which is incompatible with the narrow scope of TDM exceptions); Daniel J. Gervais, The Machine as Author, 105 Iowa Law Review 2053, 2058-59 (2020) (arguing that generative AI training internalizes expressive content, not just semantic information, and thus implicates reproduction rights and copyright incentives); Jonathan Pukas, KI-Trainingsdaten und erweiterte kollektive Lizenzen: Generierung von Werken als KI-Trainingsdaten auf Basis erweiterter kollektiver Lizenzen, cit. (stressing that TDM was meant for extracting information, not for training black-box neural networks where the content is internalized without explicit informational output).

[133] See Brauneis, Copyright and the Training of Human Authors and Generative Machines, cit. at 27-29.

retrieve and integrate external data at the inference stage, often using APIs or real-time queries. According to recent EU research, this distinction may carry significant legal weight: RAG's retrieval-based design may – in some cases – align more closely with the conditions for the TDM or temporary reproduction exceptions, especially where data is not stored persistently.[134] Moreover, licensing practices in the RAG context typically differ from those related to training corpora, reflecting stakeholders' recognition of this legal and technical divergence.[135] Failing to distinguish these models could risk blurring the boundaries between amaterially different uses of copyright-protected content. Furthermore, even when no persistent storage occurs, **the outputs of RAG systems may themselves give rise to liability**, particularly under the reproduction right or the communication to the public right—such as when summarised or excerpted content substitutes access to protected sources. These risks underscore the importance of distinguishing architectural models not only at the ingestion stage but also with regard to their **generation dynamics and downstream legal implications**.

That said, legal ambiguity remains as to whether all commercial AI training activities necessarily exceed Article 4's remit. Some Member States and scholars contend that, unless a valid opt-out has been duly exercised, the plain text of Article 4 may still support certain uses—particularly where the content has been lawfully accessed and no machine-readable reservation is present.

In order to illustrate the tension, Dornis references Google's "Smart Reply" function and the training of Stable Diffusion.[136] These systems do not merely analyse existing emails or images to extract facts or trends; they emulate styles—an internalisation of creative form that, as discussed earlier, falls outside the analytical use contemplated by the TDM exception. From a technological perspective, this difference is significant. Classic TDM might involve scanning a thousand medical articles to find correlations between drug types and side effects. Generative AI training involves processing those articles so that the system can later compute a text output that imitates their style. The former extracts knowledge; the latter reconstructs linguistic form. In the absence of a dedicated legal framework for AI training, Article 4 currently operates as the principal legal mechanism enabling such uses in many Member States—albeit imperfectly. Developers often proceed under the assumption that their activities fall within the scope of the TDM exception, particularly when content is lawfully accessed and no valid opt-out is detected. However, despite this widespread perception, empirical evidence suggests that many industry actors are in fact reluctant to rely on Article 4. The lack of legal clarity—especially concerning the effectiveness of opt-outs and the notion of "lawful access"—has prompted several major players to pursue retroactive licensing agreements or to bypass the European framework

[134] See EUIPO, The Development of Generative Artificial Intelligence from a Copyright Perspective, cit. at 272–275 (noting that while RAG differs from standard model training in both legal and technical terms, certain implementations—particularly dynamic, transient data retrieval—may align more closely with the conditions of the TDM or temporary reproduction exceptions).

[135] Ibidem.

[136] Tim Dornis, The Training of Generative AI Is Not Text and Data Mining, cit. at 70 (noting that the success of Google's "Smart Reply" system was achieved only after its training was expanded to include the BookCorpus dataset—over 11,000 novels rich in stylistic and syntactic features—suggesting that it was not semantic content alone but expressive, potentially copyright-protected elements that enabled the model to generate human-like responses).

altogether.[137] Beyond practical limitations, the opt-out mechanism raises more fundamental doctrinal concerns. Under the Berne Convention, the enjoyment and exercise of copyright shall not be subject to any formality. A system that places the burden on authors to actively reserve their rights—using machine-readable opt-outs or technical protocols—risks conflicting with this foundational principle of international copyright law. Moreover, the current opt-out regime presupposes a level of technical literacy, awareness, and infrastructural capacity that many small creators do not possess. In the absence of a collective licensing infrastructure or default opt-in rule, the mechanism fails to offer meaningful protection at scale and may disproportionately benefit large platforms that can ingest content by default unless formally excluded. As such, the opt-out does not serve as an adequate safeguard, either legally or practically. This pattern of non-reliance reinforces the study's central thesis: far from delivering genuine legal certainty, Article 4 creates an appearance of legal clarity that may not hold up under scrutiny. The result is a regulatory vacuum in which innovation proceeds without a coherent legal foundation, leaving rightsholders uncompensated and obligations ill-defined.

This misunderstanding also undermines the policy rationale for applying TDM exceptions to AI training. The TDM exception exists to support data-driven innovation and scientific research, not to enable the wholesale use of creative content without consent or compensation. As convincingly argued, this is not just a case of stretching an exception—it is a misapplication of the legal concept in its entirety.[138] EU innovation policy is not a one-dimensional pursuit of technological advancement but is grounded in a regulatory framework that balances multiple interests, including the rights of creators, the need for legal certainty, and broader public access to knowledge. Framing unlicensed large-scale ingestion of protected works as "innovation" risks distorting this balance. The type of innovation supported by Article 4 of the CDSM Directive is analytical in nature—intended to promote research and information extraction—not the commercial synthesis of expressive content. When generative models systematically absorb and recombine protected expression, they move beyond the scope of legitimate data analysis and enter a domain that raises concerns about appropriation without authorisation or oversight.

Furthermore, the risks of misclassifying AI training as TDM are not abstract. If accepted, this reading could allow AI developers to bypass licensing entirely, using massive amounts of protected works under the assumption that no rights are being infringed. This undermines the economic rights of creators and

[137] See e.g. Matt O'Brien, ChatGPT-maker OpenAI signs deal with AP to license news stories, The Associated Press (July 13, 2023). Available at https://apnews.com/article/openai-chatgpt-associated-press-ap-f86f84c5bcc2f3b98074b38521f5f75a; Helen Coster, Global news publisher Axel Springer partners with OpenAI in landmark deal, Reuters (December 13, 2023). Available at https://www.reuters.com/business/media-telecom/global-news-publisher-axel-springer-partners-with-openai-landmark-deal-2023-12-13/; Pascale Davies, OpenAI partners with European media giants in France and Spain to use content for training, Euronews (March 14, 2024). Available at https://www.euronews.com/next/2024/03/14/openai-partners-with-european-media-giants-in-france-and-spain-to-use-content-for-training; OpenAI signs multi-year content partnership with Condé Nast, The Guardian (August 20, 2024). Available at https://www.theguardian.com/technology/article/2024/aug/20/conde-nast-open-ai-deal; See Shutterstock Expands Partnership with OpenAI, Signs New Six-Year Agreement to Provide High-Quality Training Data, Press release (July 11, 2023). Available at https://investor.shutterstock.com/news-releases/news-release-details/shutterstock-expands-partnership-openai-signs-new-six-year

[138] See Tim Dornis, Generative AI, Reproductions Inside the Model, and the Making Available to the Public cit.

threatens to erode the copyright system's foundational incentive structure. Worse, it creates a legal grey zone where AI models can be trained on entire libraries of artistic and literary works under the pretext of "data mining," without any meaningful oversight or remuneration.

The legal, technical, and doctrinal evidence points to a clear conclusion: generative AI training does not fall within the scope of Articles 3 and 4 of the CDSM Directive. It exceeds the analytical purpose, violates the proportionality limits of the three-step test, and triggers reproduction rights that cannot be bypassed by invoking TDM.

As a consequence, according to a growing body of opinion, the training of generative AI models does not fall within the scope of the TDM exceptions under Articles 3 and 4 of the CDSM Directive.[139] The technical processes involved, the legal concept of reproduction, and the normative purpose of copyright protection all converge toward one conclusion: this is not mining—it is making.[140] The growing consensus now recognises that we need new legal tools and categories to address this unprecedented challenge, not a retrofit of provisions drafted for a different technological landscape.

Even supposing that the transparency obligations introduced by the EU Artificial Intelligence Act could contribute to greater oversight of training practices, such obligations remain limited in scope. Indeed, the AI Act mandates only the disclosure of summary information—not specific datasets—and does not provide mechanisms for opt-out enforcement, real-time monitoring, or model-specific auditability (see Section 2.1.3 and 2.5). Transparency, in this context, does not constitute legal authorisation. The AI Act cannot retroactively validate uses that infringe reproduction rights, nor can it substitute for compliance with copyright licensing requirements

This insight sets the stage for the following sections, where we explore how member states are implementing the current rules, how rightsholders are responding, and how emerging EU legislation like the AI Act seeks to introduce greater transparency and control in this rapidly evolving field.

Summary box

[Note: TDM exceptions are narrowly defined by Article 2(2) and constrained by systemic safeguards—most notably, the three-step test under Article 5(5) InfoSoc and the reproduction right in Article 2. These legal boundaries form the baseline for assessing the lawfulness of AI training.]

Table 3: Summary box

Legal Rationale	**Explanation**
1. Functional Limits of TDM	Article 2(2) narrowly defines TDM as analytical—not generative. Articles 3 and 4 allow reproduction or extraction only for analysis, not expressive synthesis.

[139] See supra note 76.

[140] See Tim Dornis, Generative AI, Reproductions Inside the Model, and the Making Available to the Public cit (noting that AI training is not a case of mining but one of making—requiring a fundamentally different legal treatment.)

2. Qualitative Divergence	GenAI systems recombine and simulate expressive content, moving beyond the pattern extraction that defines lawful TDM.
3. Innovation Clauses Misread	Recitals 2 and 5 encourage innovation, but only within copyright's structural limits. Article 4(3) confirms rightholders' right to control reuse via opt-outs.
4. Breach of the Three-Step Test	GenAI training is not a special case, displaces licensing markets, and prejudices authors—failing the Article 5(5) InfoSoc and Berne/TRIPS test.
5. Embedded Expression in Model Weights	Models encode elements of protected expression during training, potentially triggering reproduction rights and requiring legal accountability.
6. Reproduction Right and Transient Copies	As confirmed in *Infopaq* and *SAS Institute*, even temporary or non-visible reproductions may infringe Article 2 InfoSoc—relevant for internalised model weights and training copies
7. Distinction Between Training and RAG	Unlike model training, RAG systems retrieve external data at the inference stage and may fall under temporary reproduction or TDM exceptions—depending on data persistence and licensing context.
Conclusion	Framing generative AI training as text and data mining distorts the structure and purpose of the CDSM Directive. Such use cases fall outside the intended legal scope of Articles 3 and 4, and their inclusion would undermine rightholders' protections and violate international copyright norms..

2.1.3. Unauthorised Training and Its Legal Consequences

In light of the preceding analysis, it appears that relying on Article 4 of the CDSM Directive to justify the training of generative AI systems lacks a clear legal foundation. Training generative models involves the large-scale reproduction and internalisation of expressive content—not merely the extraction of factual information—and thus likely exceeds the definitional scope and normative intent of the TDM exception under Article 2(2) of the Directive.[141] In technical terms, training generative models involves translating expressive works into multi-dimensional vector representations that encode the stylistic, structural, and compositional features of the input data. These representations, stored in the model's weights, are not human-readable but are functionally equivalent to compressed reproductions that

[141] See Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market, art. 2(2), 2019 O.J. (L 130) 92.

enable the model to regenerate protected content. Under EU copyright law, reproduction does not require human readability or pixel-perfect duplication. It is sufficient that the act enables subsequent outputs to exploit the expressive content of the original work. This is consistent with the CJEU's technology-neutral understanding of reproduction in Infopaq and Pelham, where even transient or machine-dependent copies were found to trigger Article 2 rights.[142] Therefore, internalisation in vector space cannot be dismissed as merely analytical—it is part of the same protected act of reproduction that underpins the training process.

Given this, the internalisation of expressive works during training—though machine-mediated—cannot be dissociated from the protected act of reproduction. This distinction underscores why such training cannot qualify as a permissible analytical technique under Article 2(2) of the CDSM Directive. As clarified in Recital 8 and Article 2(2), the notion of 'analytical technique' presupposes an extraction of information, not the transformation of expression into latent vectorised form.

From both a legal and technical standpoint, these practices are not acts of analysis, but acts of reproduction[143]—and, absent a valid exception or licence, they constitute copyright infringement and may give rise to liability for damages under EU and national law. Given that many generative AI models have already been trained using protected content without consent or remuneration, this raises urgent questions of ex post liability and appropriate remedies. While Article 4 of the CDSM Directive was never designed to authorise such uses, developers have often proceeded under expansive and contested interpretations of its scope. Where no valid opt-out was respected, content may have been lawfully accessed but still unlawfully reused, and rightsholders may still be entitled to compensation—especially where outputs exhibit memorised or stylistically replicable features of protected works.

In the absence of a dedicated statutory framework, determining fair compensation will be challenging. Courts and regulators may need to consider proxies such as licensing benchmarks, dataset composition, output substitutability, or measurable economic harm to creators.[144] Legal clarity is also

---

[142] See Case C-5/08, *Infopaq International A/S v Danske Dagblades Forening*, EU:C:2009:465, §§ 33–48; and Case C-476/17, *Pelham GmbH v Hütter*, EU:C:2019:624, §§ 56–63 (illustrating the Court's technology-neutral approach: the mode of reproduction—manual, digital, or automated—is irrelevant; what matters is whether the reproduced content reflects protected expression).

[143] See e.g. U.S. Copyright Office, *Copyright and Artificial Intelligence, Part 3: Generative AI Training* (Pre-Publication Version, May 2025), at 28, available at https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf (emphasising that the ingestion and dataset creation from protected works "clearly implicate[s] the right of reproduction," making such acts presumptively infringing absent a valid exception or defence); Tim Dornis, *The Training of Generative AI Is Not Text and Data Mining*, 47 European Intellectual Property Review 65–78 (2025) (arguing that generative AI encodes and structurally replicates expression, not merely extracting patterns); Tim Dornis, *Generative AI, Reproductions Inside the Model, and the Making Available to the Public*, IIC – International Review of Intellectual Property and Competition Law (2025) (examining how internal model representations can embed protected expression, triggering reproduction and communication rights); Matthew Sag and Peter K. Yu, The Globalization of Copyright Exceptions for AI Training, 74 Emory Law Journal, (2025) (recognizing that generative AI reproduces vast volumes of copyrighted material and that its capabilities surpass those of traditional mining (e.g., extracting facts).

[144] See EUIPO, The Development of Generative Artificial Intelligence from a Copyright Perspective cit. at 14, (noting that the evolution of licensing markets for training data may be influenced by "the development of benchmark market rates," as well as output-based licensing metrics and sector-specific norms. So, these evolving practices may serve as proxies for courts and regulators assessing damages in the absence of a statutory framework).

required on whether remedies should include retrospective deletion of datasets, model retraining, or revenue-sharing mechanisms in cases of unlawful ingestion. These are not merely technical questions—they go to the heart of the EU's commitment to a rules-based, equitable copyright system.

This concern is not hypothetical. Platforms may scrape or ingest content—including private drafts, unpublished songs, or incomplete works—stored by creators on digital platforms or cloud services. Under well-established copyright doctrine, such content is protected from the moment of creation, regardless of publication.[145] The unauthorised use of these materials infringes exclusive rights under Article 2 of the InfoSoc Directive and raises serious concerns about consent, digital autonomy, and the responsibilities of platform intermediaries. It may be argued that Article 53(1)(c)–(d) of the AI Act, alongside Recital 105, presupposes that the TDM exception in Article 4 of the CDSM Directive applies to the training of general-purpose AI models. However, this interpretation overstates the legal effect of these provisions. **Article 53 is procedural in nature**: it imposes transparency and compliance obligations on providers of general-purpose AI models, but **it does not confer new rights or extend the substantive scope of copyright exceptions under EU law**. Indeed, the AI Act explicitly states that these obligations are "without prejudice" to applicable Union or national law.[146] These provisions presuppose—but do not affirm—the lawfulness of TDM-based AI training. Thus, Article 53 requires compliance if and only if the TDM exception is validly relied upon; it does not adjudicate or legitimise that reliance.

Similarly, Recital 105 acknowledges that text and data mining techniques "may be used extensively" in the context of AI training, but it merely describes current technical practices without clarifying their lawfulness.[147] While the recital reiterates the need for rightsholders authorisation where rights have been reserved, it remains silent on whether generative AI training, as a matter of law, falls within the definition of TDM in Article 2(2) of the CDSM Directive. Together, Article 53 and Recital 105 reflect a policy assumption that AI developers will rely on existing copyright exceptions—but they do not settle the legal question of whether those exceptions, as currently drafted, are applicable to generative AI training.

This ambiguity reinforces the need for doctrinal clarification or legislative reform. The AI Act presupposes legal clarity rather than establishing it—placing responsibility back on EU copyright law to

[145] See, e.g., Art. 5(2) of the Berne Convention (Berne Convention for the Protection of Literary and Artistic Works, 9 September 1886, as amended 28 September 1979, S. Treaty Doc. No. 99-27, 1161 U.N.T.S. 3), which provides that "the enjoyment and the exercise of these rights shall not be subject to any formality"; protection arises automatically and is independent of publication or registration; U.S. Copyright Off., Circular 1: Copyright Basics 1 (2021), §202. Available at https://www.copyright.gov/circs/circ01.pdf; Sam Ricketson & Jane Ginsburg, International Copyright and Neighbouring Rights: The Berne Convention and Beyond, 3rd ed. (Oxford University Press, 2022), at 236–247 (explaining that under Article 3(1) of the Berne Convention, copyright protection applies from the moment of creation, regardless of publication or formalities).

[146] See Recital 137 and Article 3(7) of Regulation (EU) 2024/1689 (AI Act), confirming that the obligations established under the AI Act are without prejudice to applicable Union or national copyright laws.

[147] Recital 105 of the AI Act describes the technical use of text and data mining techniques in AI training but stops short of affirming their legality under copyright law. It explicitly states that such use "requires the authorisation of the rightsholder concerned unless relevant copyright exceptions and limitations apply."

determine whether and to what extent generative AI training can be lawfully conducted without express authorisation or remuneration.

These legal and ethical concerns are also echoed in civil society. In 2024 and 2025, a growing number of initiatives by authors, artists, and performers—including open letters and petitions to EU institutions—have called for an immediate halt to the unlicensed use of creative works in AI training.[148] These movements reflect a widespread perception that the current interpretation of the TDM exception is being distorted to serve the interests of AI developers and platforms, at the expense of fundamental creator rights.

Any future reform of EU copyright law must reject this trajectory. Rather than accommodating large-scale ingestion under misapplied exceptions, legislative reform should reaffirm the primacy of authorial control and ensure that AI training is subject to prior consent, negotiated licensing, and fair remuneration. Exceptions must not become de facto authorisations for commercial exploitation. Instead, they must respect the constitutional balance between innovation and the protection of creative labour that lies at the core of the European copyright acquis, a balance consistently upheld by the CJEU.[149]

2.1.4. Beyond TDM: Structural Gaps in the CDSM Directive Framework

While the TDM exceptions under Articles 3 and 4 of the CDSM Directive have been interpreted as the primary legal tools enabling AI developers to access and analyse copyrighted material, it is increasingly clear that these provisions were not designed with the scale, purpose, or economic impact of generative AI systems in mind. This section outlines four essential limitations of the current TDM framework in the context of AI training: i) the narrow scope of Article 3, ii) the flaws of the opt-out mechanism under Article 4, iii) the mismatch between AI training processes and TDM objectives, and iv) the absence of compensation mechanisms for rightsholders.

---

[148] See e.g. Joint Letter to Members of the European Parliament on the Impact of Artificial Intelligence on the European Creative Community (23 July 2024), available at https://composeralliance.org/media/1651-joint-letter-to-members-of-the-european-parliament-on-the-impact-of-artific.pdf signed by major European creators' associations, calling for an end to the unlicensed use of protected works in AI training, greater enforcement of authorial consent, and the reform of Article 4 of the CDSM Directive to safeguard creator rights; Creators for Europe United, Open Letter to the European Commission for Fair, Transparent, and Legally Compliant AI Development (25 April 2025), available at https://creators-for-europe-united.eu (highlighting creators' demands for consent, transparency, and fair remuneration in AI training); Open Letter to the Attention of Ministers of Culture Ahead of the Education, Youth, Culture and Sport Council on 12–13 May 2025 (6 May 2025), available at: https://composeralliance.org/media/1864-open-letter-to-the-attention-of-ministers-of-culture-ahead-of-the-education.pdf (endorsed by a broad coalition of organisations representing writers, translators, journalists, performers, composers, visual artists, and screen directors, calling for strong safeguards for copyright and transparency under the AI Act and condemning the unauthorised use of members' works and data for AI training without consent or remuneration).

[149] See, e.g., Case C-516/17, Spiegel Online GmbH v. Beck, ECLI:EU:C:2019:625, Judgment of 29 July 2019 (clarifying that intellectual property rights are not absolute and must be balanced against other fundamental rights under the EU Charter); Case C-201/13, Deckmyn v. Vandersteen, ECLI:EU:C:2014:2132, Judgment of 3 September 2014 (stressing that copyright exceptions must be interpreted in light of the need to safeguard a fair balance between the rights and interests of authors and users); Case C-476/17, Pelham GmbH v. Hütter, ECLI:EU:C:2019:624, Judgment of 29 July 2019 (affirming that copyright exceptions must be interpreted strictly and cannot justify acts that conflict with the normal exploitation of the work).

2.1.4.1. Limits of Article 3 CDSM – TDM for Scientific Research

Article 3 of the CDSM Directive provides a targeted exception that allows research organisations and cultural heritage institutions to carry out text and data mining for scientific research purposes, provided they have lawful access to the content. This exception is unconditional: rightsholders cannot opt out. However, it is explicitly limited to non-commercial research institutions, thereby excluding most private-sector AI developers.

This strict separation between "non-commercial" and "commercial" TDM users is increasingly viewed as outdated. As recognised in Recital 11 of the CDSM Directive, AI research is increasingly conducted within public–private partnerships, where collaborations between universities, research institutions, and private companies have become standard practice. As Margoni and Kretschmer argue, this fragmented and binary structure undermines legal clarity and innovation, particularly where the line between scientific exploration and commercial exploitation is increasingly blurred.[150]

A further legal concern arises from the potential for what a recent EUIPO study describes as "data laundering": the reuse of datasets originally compiled under the scientific research exception of Article 3 for commercial AI training under Article 4. [151] This practice reflects a growing tension within the CDSM framework, as collaborative ecosystems between public research institutions and private developers make it increasingly difficult to draw a clear line between scientific and commercial use. The ability to repurpose Article 3-compliant datasets in downstream commercial contexts—without additional licensing or remuneration—raises questions about the internal consistency of the two-tiered TDM structure. More broadly, it reinforces the concern that the TDM exceptions, while appropriate for narrow research-based analysis, may be ill-suited to regulate the complex, large-scale, and economically consequential processes involved in AI model training. Addressing this misalignment may require more clearly defined boundaries between exceptions, enhanced oversight mechanisms, and further normative guidance under the AI Act.

2.1.4.2. Critiques of Article 4 CDSM – "Commercial" TDM with Opt-Out

Article 4 of the Directive extends the TDM exception to all users, including commercial entities, so long as the works are lawfully accessible and the rightsholders has not reserved their rights. At first glance, this seems to offer a viable legal route for AI developers. However, its opt-out mechanism significantly complicates its application.

Rightsholders can opt out "in an appropriate manner," such as by using machine-readable means or contractual restrictions. Yet there is no harmonised standard or technical specification defining what constitutes an "appropriate" opt-out. As a result, rights reservation practices vary widely—from metadata tags to terms of service—and there is no unified system for detecting and enforcing them. This creates significant legal uncertainty for developers and imposes a constant compliance burden: companies must monitor every scraped or licensed source for potential opt-outs. In practice, this has

[150] See Thomas Margoni and Martin Kretschmer, A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology, 71 *GRUR International* 685-701 (2022).

[151] See EUIPO, The Development of Generative Artificial Intelligence from a Copyright Perspective (May 2025), cit. at 117 (raising concerns about dataset reuse by commercial developers through partnerships with scientific institutions).

led to criticism that Article 4 offers a "pseudo-permission" system.[152] While nominally an exception, it may be rendered ineffective in large-scale web scraping contexts, such as those used for training AI models, where checking opt-out signals across billions of pages is impractical. The result is a legal framework that neither reliably permits nor clearly prohibits AI training—leaving all stakeholders in a regulatory limbo, and prompting leading AI developers to secure bespoke content licences instead.[153] This legal fragmentation is also echoed in recent EU-commissioned research, which found that stakeholders are deeply divided on the legal sufficiency and interoperability of opt-out mechanisms, especially when AI systems ingest vast and heterogeneous datasets.[154] This critique has also been echoed beyond the EU. In particular, the U.S. Copyright Office has recently warned that opt-out mechanisms are difficult to implement fairly and effectively—especially for individual authors—and fundamentally conflict with the permissions-based logic of copyright law.[155]

2.1.4.3. Structural Limitations of the Opt-Out Mechanism under Article 4(3)

The concerns above have been further substantiated by a range of technical, legal, and policy analyses – including the recent 2025 EUIPO study – which identifies the current opt-out mechanisms under Article 4 as fragmented, technically fragile, and largely unenforceable, raising serious doubts about their effectiveness as a reliable safeguard for rights holders.[156] A closer examination reveals several critical flaws that undermine its intended function and broader fairness in the copyright ecosystem. Although Article 4 of the CDSM Directive introduces a novel opt-out mechanism intended to preserve the freedom of rightsholders to withhold their works from text and data mining by commercial actors, its practical implementation is fraught with conceptual and logistical shortcomings. In theory, rightsholders can exclude their content through location-based signals (e.g. robots.txt) or unit-level

[152] See Paul Keller, Generative AI and copyright: Convergence of opt-outs?, Kluwer Copyright Blog (November 23, 2023). Available at https://copyrightblog.kluweriplaw.com/2023/11/23/generative-ai-and-copyright-convergence-of-opt-outs/

[153] Recent press coverage illustrates the trend: Shirin Ghaffary, OpenAI has been "in talks with dozens of publishers" to licence content for model training and downstream display (Bloomberg, 4 Jan 2024). Available at https://www.bloomberg.com/news/newsletters/2024-01-04/openai-says-it-s-in-talks-with-dozens-of-publishers-about-licensing-content; Maria Deutscher, OpenAI signs content licensing agreement with the Financial Times (SiliconAngle, 29 Apr 2024). Available at https://siliconangle.com/2024/04/29/openai-signs-content-licensing-agreement-financial-times/; Mackenzie Ferguson, "Anthropic Reaches Landmark Settlement with Music Publishers Over AI-Generated Lyrics", OpenTools AI News, 24 May 2024, https://opentools.ai/news/anthropic-reaches-landmark-settlement-with-music-publishers-over-ai-generated-lyrics.

[154] See European Commission, Study on Copyright and New Technologies: Copyright Data Management and Artificial Intelligence, European Commission, 2022, at 198–201. Available at https://op.europa.eu/publication-detail/-/publication/cc293085-a4da-11ec-83e1-01aa75ed71a1

[155] See U.S. Copyright Office, Copyright and Artificial Intelligence, Part 3: Generative AI Training, cit. at 74–75, noting that "[t]he Copyright Act establishes an opt-in, permissions-based regime… There is no basis in law or policy for imposing an opt-out regime," and expressing concern that opt-out mechanisms "may raise practical and fairness concerns, especially for individual creators unfamiliar with machine-readable reservations."

[156] See e.g. European Union Intellectual Property Office, Development of Generative Artificial Intelligence from a Copyright Perspective (May 2025), at 15–17, 164–234. Available at https://www.euipo.europa.eu/en/publications/genai-from-a-copyright-perspective-2025 (describing current opt-out mechanisms under Article 4 as technically limited, fragmented across sectors, and ultimately lacking enforceability, with no single standard in place).

metadata (e.g. IPTC "noAI" tags).[157] In practice, however, both methods prove ineffective and are structurally misaligned with the realities of AI training at scale. Location-based methods only affect content hosted on platforms the rightsholder controls—leaving copies disseminated elsewhere vulnerable.[158] Unit-based tagging systems, such as those relying on embedded metadata (e.g. IPTC or C2PA), offer limited protection in practice, as metadata can be easily removed or ignored—and cannot be applied at all to certain widely used formats such as plain text, code, or scraped HTML.[159] Adoption has also been extremely limited, due to low awareness, technical barriers, and the lack of harmonised standards.[160] The system further imposes a binary choice on creators: be visible to the public or protect their content from AI—but not both.[161] Such a design ignores legitimate intermediate positions, such as permitting citation or reference without allowing training replication. As a result, the administrative

[157] See e.g. Paul Keller, Considerations for Opt-Out Compliance Policies by AI Model Developers, Open Future, May 16, 2024, https://openfuture.eu/wp-content/uploads/2024/05/240516considerations_of_opt-out_compliance_policies.pdf (distinguishing between location-based methods—such as robots.txt, ai.txt, and HTTP headers—and unit-based tools like embedded metadata (e.g. IPTC tags, C2PA), ISCC codes, or watermarking, and noting that these mechanisms remain fragmented and are adopted inconsistently across platforms and content types); Hanjo Hamann, Artificial Intelligence and the Law of Machine-Readability: A Review of Human-to-Machine Communication Protocols and their (In)Compatibility with Article 4(3) of the Copyright DSM Directive, 15 JIPITEC 102-121 (2024) (observing that the proliferation of opt-out mechanisms "currently precludes any effective reservation of TDM rights," due to inconsistent implementation, doctrinal ambiguity, and technical limitations); See Ed Newton-Rex, The Insurmountable Problems with Generative AI Opt-Outs (Nov. 2024), available at: https://ed.newtonrex.com/s/The-insurmountable-problems-with-generative-AI-opt-outs.pdf (identifying fundamental limitations of opt-out mechanisms, including the ineffectiveness of location- and unit-based approaches, their low adoption rate, and their failure to provide meaningful or enforceable control over downstream uses of protected content).

[158] See e.g. Chien-Yi Chang and Xin He, The Liabilities of Robots.Txt. University of Hong Kong Faculty of Law Research Paper No. 2025/06, Available at SSRN: https://ssrn.com/abstract=5159436 (explaining that the robots.txt file only affects content hosted at the domain where the webmaster has control and explicitly noting the legal and practical limits of relying on this protocol for content protection). See also EUIPO, The Development of Generative Artificial Intelligence from a Copyright Perspective (May 2025), cit. at 228–229 (noting that location-based opt-outs such as robots.txt are limited to content hosted on controlled domains and do not apply to redistributed copies).

[159] See Hanjo Hamann, Artificial Intelligence and the Law of Machine-Readability, cit. 15 at 8, 11 (explaining that metadata-based opt-outs like IPTC and C2PA are inapplicable to formats such as plain text or HTML, and highlighting the lack of standardisation and the ineffectiveness of conflicting metadata tags for content protection). See also EUIPO, The Development of Generative Artificial Intelligence from a Copyright Perspective (May 2025), cit. at 173–175, 208 (discussing the practical limitations of embedded metadata systems, including ease of removal, lack of support for certain file types, and limited crawler compliance).

[160] See e.g. Alex Bocharov et al., Declare Your AIndependence: Block AI Bots, Scrapers and Crawlers with a Single Click, Cloudflare Blog (3 July 2024), available at: https://blog.cloudflare.com/declaring-your-aindependence-block-ai-bots-scrapers-and-crawlers-with-a-single-click (reporting that AI bots like Bytespider and GPTBot accessed over 40% and 35% of Cloudflare-protected websites, respectively, and noting widespread user demand for simple blocking tools due to the complexity and inconsistency of existing opt-out mechanisms); See Bron Maher, Revealed: which news sites are blocking the AI web crawlers, Press Gazette (27 February 2024), available at: https://pressgazette.co.uk/platforms/news-sites-block-ai-web-crawlers-chatgpt-google/ (reporting that 42.5% of major UK and US news sites had not blocked any AI bots, highlighting limited adoption and inconsistent implementation of AI crawler-blocking measures); EUIPO, The Development of Generative Artificial Intelligence from a Copyright Perspective (May 2025), cit. at 208–234 (noting that uptake of reservation tools remains low due to limited awareness, technical hurdles, and the absence of standardised, widely adopted protocols).

[161] See e.g. Ed Newton-Rex, The Insurmountable Problems with Generative AI Opt-Outs, cit.; EUIPO, The Development of Generative Artificial Intelligence from a Copyright Perspective (May 2025), cit. at 208–230 (noting that current reservation tools often require removing content from public indexing to opt out of AI use, thus forcing creators to choose between visibility and protection).

burden on creators—especially individual and small-scale authors—is considerable: they must apply opt-outs manually to each new work, often across platforms they do not control. An additional concern is that the mechanism may be seen as implicitly legitimising prior infringement: even when a rightsholder expresses an opt-out, it does not trigger any obligation to retrain models or delete previously ingested works.[162] As such, the opt-out functions only as a prospective, non-retroactive safeguard—potentially reinforcing asymmetries of access and remuneration.[163] From a systemic fairness perspective, this disproportionately harms small creators, who often lack the resources or technical capacity to monitor dataset inclusion or assert their preferences.[164] It is therefore highly improbable to design an opt-out mechanism that both achieves widespread awareness among eligible individuals and remains adaptable to the continuously evolving landscape of web crawlers and web scraping.[165]

In light of these cumulative shortcomings, **the opt-out solution** in Article 4(3) **appears not only legally ambiguous** and under-specified, **but also functionally unworkable**. A coherent and enforceable copyright framework must instead consider returning to a permissions-based "opt-in" regime—one where the default is protection, not presumed access (see Section 4.1(D)).

2.1.4.4. Mismatch with AI Training Needs

The underlying conceptual purpose of TDM—to enable the extraction of information or knowledge—is fundamentally different from the purpose of generative AI training, which involves copying and internalising expressive content at scale. The datasets used in this process typically include not only factual material, but also literary texts, visual artworks, software code, music, and other works protected by copyright.

Critics argue that this makes the CDSM TDM exceptions ill-suited to justify the creation of training datasets for generative AI systems or that were not drafted in light of GenAI.[166] Indeed, the impression

---

[162] See EUIPO, The Development of Generative Artificial Intelligence from a Copyright Perspective (May 2025), cit. at 228–234 (noting that most reservation mechanisms are not retroactive and do not require retraining or removal of previously ingested content from AI models).

[163] See e.g. OECD, Intellectual Property Issues in Artificial Intelligence Trained on Scraped Data, cit., at 20–25 (noting that most rightsholders—particularly individual creators—lack the technical means to monitor whether their works have been scraped and used for AI training, highlighting the structural imbalance and the ineffectiveness of current opt-out mechanisms.)

[164] Ibidem.

[165] Web crawling refers to the automated process of systematically browsing the web to index publicly available content, typically for search engine purposes. Web scraping, by contrast, involves the automated extraction of specific data or content from websites, often at scale and beyond indexing functions. See EUIPO, The Development of Generative Artificial Intelligence from a Copyright Perspective (May 2025), cit. at 356.

[166] See e.g. Giorgio Franceschelli and Mirco Musolesi, On the creativity of large language models. AI & Soc 1, 3 (2024) (arguing that these exceptions were not conceived with GDL in mind and point out the inadequacy of current copyright laws); João Pedro Quintais, Generative AI, copyright and the AI Act, 56 Computer Law & Security Review 1 (2025); Christophe Geiger, When the Robots (Try to) Take Over: Of Artificial Intelligence, Authors, Creativity and Copyright Protection in Florent Thouvenin et al. (eds.), Kreation Innovation Märkte – Creation Innovation Markets: Festschrift Reto M. Hilty' (2024), 67, 77 (claiming that the TDM exception was "not designed to cover machine learning by generative AI systems"); Katharina de la Durantaye, Control and Compensation. A Comparative Analysis of Copyright Exceptions for Training Generative AI, IIC - International Review of Intellectual Property and Competition Law 1-34 (2025) (highlighting the limitations and legal

is that these exceptions were not drafted with the goal of enabling systems to be trained to compute outputs that replicate the style or structure of protected content. They were intended to support data analytics, not content replication. This view is also supported by policy-oriented research highlighting the conceptual gap between TDM for information extraction and the expressive replication intrinsic to generative AI.[167]

This disconnect underscores the broader concern that the CDSM TDM provisions were tailored to a different technological paradigm. As such, they may no longer provide an adequate or reliable legal foundation for the practices that underpin the next wave of AI development.

2.1.4.5. Lack of Remuneration and Enforcement Mechanisms

Perhaps the most tangible gap in the current framework is the absence of any remuneration or compensation mechanism for rightsholders. Articles 3 and 4 of the CDSM Directive permit certain uses of protected content without triggering any obligation to pay or licence, even when that content is used systematically and at scale in commercially valuable AI products.[168]

This has generated deep concern across the creative and publishing sectors. Authors' guilds and collecting societies have called for the introduction of new compensation mechanisms, such as equitable remuneration schemes, statutory levies, or collectively managed licences.[169] The lack of such mechanism's risks replicating the systemic imbalances observed in other digital markets—where creators provide the raw materials, but intermediaries capture the economic value.[170]

These structural flaws in the CDSM framework have not been adequately addressed in the broader legislative response to AI. In fact, the AI Act incorporates copyright provisions without resolving the underlying legal and technical challenges—potentially reinforcing, rather than correcting, the current dysfunction.

---

uncertainty of the Article 4 opt-out regime and its ineffectiveness in practice); See Peter Mezei, A saviour or a dead end? Reservation of rights in the age of generative AI' 46 Eur. IP Rev. 461, 463 (2024).

[167] See European Commission, Study on Copyright and New Technologies: Copyright Data Management and Artificial Intelligence, European Commission, 2022, at 210. Available at https://op.europa.eu/publication-detail/-/publication/cc293085-a4da-11ec-83e1-01aa75ed71a1 (noting that several stakeholders advocate for limiting TDM exceptions to uses that generate information, and excluding use cases focused on generating creative output).

[168] See European Commission, Study on Copyright and New Technologies: Copyright Data Management and Artificial Intelligence, European Commission, 2022, at 228–30. Available at: https://op.europa.eu/publication-detail/-/publication/cc293085-a4da-11ec-83e1-01aa75ed71a1 (showing that a scenario where authors can oppose all AI training via moral rights results in the highest potential revenue for rightholders, albeit with negative implications for AI developers and EU innovation competitiveness).

[169] See e.g. See The Authors Guild, AI Licensing for Authors: Who Owns the Rights and What's a Fair Split? December 12, 2024. Available at https://authorsguild.org/news/ai-licensing-for-authors-who-owns-the-rights-and-whats-a-fair-split/; GEMA / SACEM joint study, AI and music: Generative Artificial Intelligence in the music sector . Available at https://www.gema.de/en/news/ai-study

[170] A notable precedent is the case of news publishers whose content was widely used by platforms such as Google and Facebook—without remuneration—until legal intervention through Article 15 of the CDSM Directive sought to address this disparity. As in that case, creators provide the raw material (journalistic or expressive works), while powerful intermediaries extract disproportionate economic value.

While the AI Act now formally includes copyright compliance through Article 53(1)(c), this provision risks being more symbolic than substantive. The recently drafted General-Purpose AI Code of Practice (third draft) illustrates this concern.[171] However, this Code is a voluntary instrument, whereas Article 53(1)(c)-(d) of the AI Act imposes binding obligations. GPAI providers are legally required to establish a copyright compliance policy and to publish dataset summaries based on a Commission-defined template. These rules, while mandatory in law, still require further technical elaboration and enforcement, which will fall under the purview of the newly established AI Office. Although both instruments seek to enhance transparency, their legal weight and scope differ significantly. The Code of Practice, despite outlining detailed commitments—such as machine-readable opt-outs, copyright policies, and complaint mechanisms—remains grounded in vague standards and voluntary adherence. As a result, it lacks any enforcement mechanism, meaning that, in the absence of legal consequences, AI developers—particularly those based outside the EU—have little real incentive to comply. As further highlighted by stakeholders such as COMMUNIA, the third draft of the Code of Practice has backtracked on key commitments regarding copyright transparency and rights reservation compliance.[172] Compared to earlier drafts, the latest version replaces mandatory disclosure measures with vague encouragements and continues to rely primarily on the outdated Robot Exclusion Protocol for opt-outs.[173] These changes have raised concerns that the Code will ultimately fail to provide meaningful safeguards for rightsholders, particularly when AI model developers remain free to interpret compliance standards and avoid public accountability.[174] Another limitation in the third draft concerns the narrow scope of commitments regarding rights reservation compliance. The Code applies these obligations only to data obtained through web crawling, thereby excluding other prevalent data acquisition methods such as dataset downloads, API harvesting, or third-party aggregations.[175] This design choice is difficult to reconcile with the broader mandate of Article 53(1)(c), which requires a general policy for copyright compliance regardless of how the training data is obtained. As recent commentary has noted, this creates an artificial distinction that may undermine enforcement and create incentives for developers to bypass compliance simply by shifting their data collection strategies.[176] Respect for rights reservation mechanisms under Article 4(3) CDSM should not depend on the method of access, but on the use of protected works for generative AI training—where copyright concerns are most acute.

---

[171] See European Commission, Working Groups of the First General-Purpose AI Code of Practice, Third Draft of the General-Purpose AI Code of Practice – Copyright Section, April 2024, available at: https://ec.europa.eu/newsroom/dae/redirection/document/113606

[172] See Teresa Nobre, 3rd Draft of the GPAI Code of Practice: Copyright Transparency Is Unwanted, and It Shows, COMMUNIA (Apr. 4, 2025), available at https://communia-association.org/2025/04/04/3rd-draft-of-the-gpai-code-of-practice/

[173] Ibidem.

[174] Ibidem.

[175] See Paul Keller, Is web scraping the only copyright concern for AI? The Code of Practice's blind spot, COMMUNIA (March 21, 2025), available at https://communia-association.org/2025/03/21/is-web-scraping-the-only-copyright-concern-for-ai-the-code-of-practices-blind-spot/

[176] Ibidem.

This institutional blind spot is compounded by persistent technical and legal flaws in the opt-out mechanism itself. Article 4(3) of the CDSM Directive, which is meant to safeguard rightsholders' interests, has proven largely ineffective in practice. As we have seen, there are still no harmonised or widely adopted technical standards for expressing reservations in a machine-readable way. Even more critically, developing a truly functional and universally applicable opt-out mechanism poses significant technical and legal challenges. In fact, much of the online content used in AI training is uploaded by third parties—not the rightsholders themselves—making it practically impossible for creators to assert their rights effectively. As a result, protected works are routinely used in AI training without true consent or remuneration, raising fundamental concerns about fairness and enforceability.

Leading collective rights organisations have also voiced concern that the TDM exceptions were never intended to legitimise the use of protected works for generative AI training.[177] In this context, the inclusion of Article 53(1)(c) in the final AI Act raises concerns about whether procedural transparency tools are being asked to compensate for deeper unresolved tensions in the copyright framework. While it has been argued that the provision reinforces existing rights—particularly through its emphasis on opt-out compliance—this effect remains contingent on effective implementation and does not resolve the normative and economic imbalances at play. Continuing with the current framework, without reassessing its legal foundations, may perpetuate a system that appears balanced but does not fully address concerns around unremunerated use of creative works. This raises questions about compliance with the principle of proportionality and the EU's broader commitment to a fair and balanced copyright regime, as reflected in Recital 3 of the InfoSoc Directive. If the goal is to support sustainable innovation and a fair digital economy, more comprehensive legislative responses must be considered.

Taken together, these critiques reveal that the current CDSM TDM exceptions:

1) Were not conceived with AI training practices in mind;

2) Fail to accommodate hybrid public–private R&D models;

3) Impose impractical burdens on users and provide weak enforcement for rights reservation;

4) And offer no financial recognition for the underlying contribution of creators.

While the AI Act introduces transparency obligations—such as the requirement for general-purpose AI developers to publish summaries of training data—these are merely disclosure tools. They do not resolve the deeper legal and economic misalignment between copyright protections and the realities of generative AI.

[177] Dr. Tobias Holzmüller, GEMA, has observed that "regardless of whether the current text and data mining (TDM) provisions are formally applicable to generative AI tools, it is important to recognise that such a technology was not in the minds of legislators when the TDM rules were originally conceived. These provisions were never designed to accommodate the use of creative works as training material for tools that generate vast amounts of new output that directly competes with human-created works." See Tobias Holzmüller, CEO of GEMA, personal communication with the author, email dated May 7, 2025.

2.1.4.6. Moral Rights as a Regulatory Pressure Point

The current copyright framework focuses predominantly on economic rights and their exceptions. However, a growing number of stakeholders argue that this approach fails to capture the full range of concerns raised by generative AI, particularly those tied to the ethical use and reputational impact of machine-generated content derived from protected works.

In this context, moral rights—especially the right of integrity—are emerging as a serious regulatory pressure point. According to a recent EU-commissioned study, 67% of surveyed stakeholders supported the view that rightsholders should be allowed to invoke moral rights to oppose the use of their works in AI training, even where economic rights-based exceptions, such as those under Article 3 or 4 of the CDSM Directive, would otherwise apply.[178]

This signals a paradigm shift: authors and creators are not only concerned with economic exploitation but also with the symbolic, ethical, and reputational consequences of having their works used in opaque and potentially distorting AI systems. The lack of harmonisation of moral rights across the EU further exacerbates this challenge, creating uncertainty about whether such rights can be relied upon effectively to restrict or contest AI training practices.

From a policy perspective, this trend suggests that future legal reforms may need to go beyond questions of remuneration and opt-out logistics, to explicitly address the dignitary dimensions of authorship in the age of algorithmic content generation. As the next sections will explore, more coordinated legal reform is needed to ensure that AI development respects creator rights while enabling innovation in a fair and transparent way.

2.1.5. Anticipating the CJEU's Ruling in Case C-250/25

The doctrinal and policy concerns discussed thus far—particularly regarding the scope of the TDM exception in Article 4 CDSM and the reproduction of protected content by generative AI systems—are no longer purely academic. In a significant development, the Court of Justice of the European Union (CJEU) has been called upon to interpret precisely these issues. The preliminary ruling request in Like Company v. Google Ireland (Case C-250/25) presents the first opportunity for the Court to clarify whether, and under what conditions, AI training and AI-generated outputs implicate copyright and related rights under EU law.[179] However, while the referral raises questions about AI training, the factual background of the case suggests that the disputed output was generated in response to user prompts by accessing live web content (potentially via Retrieval-Augmented Generation),[180] rather than from

[178] See European Commission, Study on Copyright and New Technologies: Copyright Data Management and Artificial Intelligence, European Commission, 2022, at 230. Available at: https://op.europa.eu/publication-detail/-/publication/cc293085-a4da-11ec-83e1-01aa75ed71a1

[179] See CJEU, Case C-250/25, Like Company v. Google Ireland, preliminary reference lodged on 3 April 2025. Referral from Fővárosi Törvényszék (Budapest Metropolitan Court), Hungary. Available at https://curia.europa.eu/juris/liste.jsf?num=C-250/25&language=en

[180] Retrieval-Augmented Generation (RAG) is a method in which a generative AI system supplements its internal knowledge by retrieving relevant documents or data from external sources—like wikis, databases, or web pages—at the time of a user query. This retrieved information is then incorporated into the prompt for content generation. RAG allows the system to provide up-to-date or context-specific responses without having been trained directly on the referenced material. See

training data.[181] This raises doubts about whether the alleged infringement is genuinely related to the training process—and, consequently, whether the CJEU will engage with the training-related questions in its ruling. This case thus potentially offers a first concrete and timely test of the legal framework analysed in this study, and its outcome may shape both future jurisprudence and the trajectory of legislative reform.

The case raises multiple interpretive questions under both the InfoSoc Directive and the CDSM Directive. A press publisher alleges that Google's LLM reproduced parts of a newspaper article in chatbot answers without permission. In its referral, the Budapest court highlights four pivotal issues: (i) whether a chatbot's verbatim display of protected press content in response to user queries constitutes an act of **communication to the public**, and whether the **predictive nature** of LLM responses affects that qualification; (ii) whether the act of **training** a generative AI system on such content constitutes **reproduction** within the meaning of EU copyright law; (iii) if so, whether such reproduction falls within the **text-and-data mining (TDM) exception** under Article 4 of the CDSM Directive; and (iv) whether reproducing or displaying protected content in chatbot responses, based on **user prompts**, constitutes a further **act of reproduction** attributable to the AI service provider.

These questions strike at the core of the arguments presented throughout this study, particularly concerning the misapplication of the TDM exception to generative AI systems and the output-side risks associated with expressive reconstruction. The answers provided by the Court will probably help determine whether current law adequately balances innovative machine-learning uses against the rights and revenues of authors and publishers, or whether legislative intervention is needed to restore that balance.

Although one must be cautious in predicting judicial outcomes, the Court's past jurisprudence and the framing of the referral provide useful indicators. It is plausible that the CJEU will adopt a rights-protective stance, in line with the EU's overarching commitment to strong copyright enforcement. **On the first question**—whether a chatbot's output of protected text constitutes an act of reproduction and making available—the likely answer is affirmative. If the facts establish that the output incorporated expressive elements of the newspaper content (beyond insubstantial fragments), the Court can be expected to affirm that both the author's rights and the press publisher's related right are implicated. The mere involvement of AI prediction does not alter the legal characterisation of the act, so long as what the end-user receives is essentially protected expression originating from the claimant's work. Moreover, the fact that such content is generated in response to a user prompt does not necessarily shift responsibility to the end-user. The Court may clarify that the reproduction is attributable to the

Patrick Lewis et al., Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, in 12 Advances in Neural Info. Processing Sys. 33, at 9459, 9460, available at https://arxiv.org/abs/2005.11401; Kim Martineau, What Is Retrieval-Augmented Generation?, IBM (Aug. 22, 2023), https://research.ibm.com/blog/retrieval-augmented-generation-RAG

[181] For a more detailed discussion on Copyright Implications of RAG, see EUIPO, The Development of Generative Artificial Intelligence from a Copyright Perspective (May 2025), cit. at 273 et seq. (noting that RAG applications, which rely on retrieving vectorised embeddings from external databases during inference, differ from conventional AI training and may not fall clearly under the TDM exceptions in Articles 3 and 4 of the CDSM Directive; highlighting that the prevailing practice in commercial RAG systems is direct licensing rather than reliance on copyright exceptions, particularly due to concerns over the duration and purpose of reproductions involved in static RAG implementations).

AI service provider, given that the system's architecture and training choices determine the scope and fidelity of the output. If the chatbot reliably generates protected material upon prompt, the legal implication is that the provider enables—or may be held liable for—the unauthorised reproduction of copyrighted content. It is therefore reasonable to anticipate that the CJEU will reiterate that allowing such outputs without authorisation would undermine the high level of protection guaranteed by the InfoSoc and CDSM Directives. In practical terms, a ruling along these lines would put AI developers on notice that they risk infringement if their models output non-trivial portions of copyrighted text. It would also empower rightsholders—especially press publishers—to demand compliance, either through licensing arrangements or through injunctive relief and damages if unlicensed reproductions persist.

On the second question—whether AI training implicates the reproduction right—the Court is again likely to respond affirmatively. Given that training involves copying vast quantities of data into memory and embedding information in model parameters, it squarely fits the broad definition of reproduction: any direct or indirect, temporary or permanent copying by any means. While the CJEU has not yet ruled specifically on TDM, the very creation of Articles 3 and 4 of the CDSM Directive suggests that, in the EU's legal framework, acts of copying for data analysis qualify as restricted acts—permissible only where a specific exception or limitation applies. The Court may therefore emphasise that Europe proactively enacted Articles 3 and 4 of the CDSM Directive precisely because, absent these provisions, even automated analysis could violate reproduction rights. Thus, insofar as generative AI training exceeds what those exceptions permit, it remains subject to the default rule: no reproduction of protected material without permission.

We may see the Court draw a principled distinction between merely reading or observing works—which is not a restricted act—and making digital copies of those works—which is. Training a generative AI system undeniably involves the latter. Unless the activity falls squarely within a narrowly construed exception, it triggers the author's exclusive rights.

The third question—the applicability of Article 4's TDM exception—is arguably the most complex and decisive. Here, the Court's answer is likely to be more nuanced. It will presumably examine the cumulative conditions of Article 4: lawful access, no reservation by the rightholder, and that the copies are made solely for TDM purposes. Whether generative AI training meets those criteria will be central to the ruling. One possible outcome is that the CJEU holds that Article 4 can, in principle, apply to the acts of reproduction during AI training—provided the Member State has properly implemented the exception and the rightsholder did not opt out. However, the Court may also clarify that Article 4 does not extend to any subsequent use of the content, such as delivering excerpts to the public. In other words, the exception might protect the input stage (data copying), but not the output stage (public dissemination).

The Hungarian court's phrasing already distinguishes between these phases, which may lead the CJEU to rule that—even if the initial data ingestion was covered by Article 4—the chatbot's output of protected text remains an infringing act that falls outside the exception. Such a conclusion would mean that Google could not avoid liability for its chatbot's responses even if the training data was lawfully mined.

Another possibility—one more closely aligned with this study's analysis—is that the CJEU will implicitly or explicitly **narrow the scope of Article 4 in the context of generative AI**. The Court could stress that Article 4's underlying purpose is to enable knowledge extraction, not to facilitate the creation of substitute content. Interpreting it to permit generative uses would risk upsetting the fair balance of rights. The Court may also invoke Recital 9 of the CDSM Directive, which underscores the need to safeguard the legitimate interests of rightholders, to caution against overly expansive readings.

The most emphatic outcome, though less certain, would be for the Court to indicate that **generative AI training does not qualify as "TDM" within the intended meaning of Article 4 at all**—effectively endorsing the view that such activity lies entirely outside the exception's ambit. This would close the door on unlicensed training where no opt-out has been made, requiring AI developers to seek explicit permission in all such cases.

Whichever way the judgment ultimately falls, the **policy consequences will be considerable**. A pro-rightsholders ruling—affirming infringement and limiting Article 4's scope—would vindicate calls for reform. It would underscore that the current framework was not designed with generative AI in mind and that relying on a fragmented opt-out regime is unsustainable. Policymakers should seize on such a decision to advance the comprehensive changes this study advocates: converting the TDM regime into an opt-in system; establishing collective licensing or remuneration schemes; and enhancing transparency and institutional oversight so that rights can be effectively managed in the AI context.
In practical terms, future legislation could replace Article 4's exception with a requirement that AI developers obtain licences—potentially through collective bodies—for any large-scale training on protected content. At the same time, a statutory remuneration right could ensure that, even where direct licensing is impractical, rightsholders are compensated for the use of their works. The creation of an **AI & Copyright Unit**, or a similar oversight body (see Section 4), could be fast-tracked to supervise these obligations and mediate between AI firms and the creative sector.

From a regulatory perspective, the policy response must also address output-side concerns. If the CJEU rules that outputting protected content is unlawful, regulators should consider technical standards or regulatory guidelines to prevent such leakage—for example, requiring large-scale AI models to implement content filters or "copy-detection" systems. Moreover, clarifying the relationship between the press publishers' right and AI would also be warranted, potentially by amending Article 15's recitals or enforcement mechanisms to explicitly include AI-generated news summaries.

If, on the other hand, the CJEU were to adopt a more permissive interpretation—for instance, finding that Article 4 applies to generative training as long as there is no opt-out—the need for legislative reform would become even more urgent. A broad reading of the TDM exception that effectively legitimises uncompensated use of vast volumes of protected works would alarm many authors and publishers. Parliament would then need to intervene decisively to recalibrate the law, lest the core principles of copyright be undermined. In that scenario, one could expect pressure to amend the Directive or introduce new provisions that clearly exclude generative AI training from the scope of Article 4 or impose remuneration obligations even where the exception applies.

In any case, whether the Court adopts a narrow or broad stance, **Case C-250/25 will signal that EU copyright law requires fine-tuning in the AI era**. Lawmakers must be ready to act on that signal. They must ensure that legal coherence is restored: creators should not be left uncompensated or without remedy simply because a user accesses their work via a chatbot rather than through a traditional interface.

Anticipating the CJEU's intervention in this case leads to a common endpoint: the recognition that our current legal toolkit is under strain and must be updated. The analysis in this study has already pointed toward the necessary direction of reform: it calls for clear rules on input/output distinctions, harmonised opt-out (or opt-in) mechanisms, transparency obligations, and equitable licensing models, so that innovation can flourish without hollowing out authors' rights. A CJEU ruling in Like Company will likely reinforce these points—either by confirming that generative AI uses are not exempt and must operate within a new licensing/remuneration framework, or by exposing gaps that policymakers will then urgently need to address. The European Parliament, as the driver of policy reform, should treat the forthcoming judgment as a catalyst. By proactively legislating in line with the principles of fairness, transparency, and accountability outlined in this study, lawmakers can help reinforce the adaptability and integrity of Europe's copyright system. The ultimate objective is a balanced regime where generative AI can develop responsibly—training on data with permission and/or compensation, and producing outputs with proper regard for others' rights—thus preserving the incentive to create and the diversity of cultural and news content on which both AI and democratic society depend. In doing so, the EU can ensure that generative AI serves the public interest without eroding the foundations of authorship, creativity, and informational pluralism.

2.1.6. Comparative Jurisdictional Approaches to TDM: Lessons for EU Policy Reform

A detailed comparative analysis of the regulatory frameworks governing text and data mining (TDM) in leading jurisdictions reveals significant divergences in legal philosophy, scope, and operational clarity—each offering instructive, though context-specific, insights for the EU's ongoing reassessment of its copyright exceptions under the CDSM Directive. Japan, the United Kingdom, and the United States provide notably distinct approaches, each with distinctive advantages and underlying policy trade-offs that the European legislator should consider in future reforms.

Japan reflects a notably permissive model for TDM activities. The 2018 amendment to the Japanese Copyright Act introduced Article 30-4, a broad exception permitting the use of copyrighted works for purposes that do not involve the enjoyment of the expressive content of the work.[182] This "non-enjoyment" standard marks a conceptual shift: the act of using a work for computational purposes—

[182] Act No 30 of 25 May 2018. See in detail Japan Copyright Office (JCO), 'Outline of the Amendments to the Copyright Act in 2018' (2019) 4 Patents & Licensing 10. For a more detailed comment, see Tatsuhiro Ueno, The Flexible Copyright Exception for 'Non-Enjoyment' Purposes - Recent Amendment in Japan and Its Implication Get access, 70 GRUR International 145-152 (2021).

such as machine learning or statistical analysis—is not considered copyright infringement, provided it does not aim to reproduce the author's expression as such.[183]

While the Japanese model is often cited for its conceptual clarity and technological neutrality—it applies across all types of copyrighted works and users[184]—it also raises questions about the scope of permissible reuse. For instance, the law does not impose explicit limits on the retention or dissemination of TDM corpora, nor does it fully clarify how this interacts with contractual restrictions or downstream uses potentially involving partial reconstitution of expressive features.[185] Accordingly, the precise scope and implications of the Japanese exception remain the subject of ongoing legal and policy debate.[186]

The justification for this approach lies in a doctrinal distinction: copyright is only infringed when a work is used "as a work"—that is, in a way that communicates expressive elements to human users. Japan's framework reflects a policy orientation that prioritises innovation. However, whether this approach aligns with the EU's more cautious stance remains open to debate.[187] While it has been welcomed in some scholarly circles,[188] doubts persist as to whether such a model can be reconciled with the EU's copyright principles, particularly in the context of AI development. As such, although the Japanese system offers a valuable comparative lens, it cannot be regarded as a ready-made regulatory blueprint.

The United Kingdom, although no longer subject to EU copyright directives, presents a more narrowly framed yet instructive model. In 2014, the UK Copyright, Designs and Patents Act (CDPA) was amended to include Section 29A, which permits the copying of works for text and data analysis, provided the use is for non-commercial research and the user has lawful access.[189] This provision was introduced independently, following the recommendations of the Hargreaves Review, as part of the UK's broader copyright reform agenda.[190] Access must be lawful, typically meaning that researchers

---

[183] See Tatsuhiro Ueno, The Flexible Copyright Exception for 'Non-Enjoyment' Purposes - Recent Amendment in Japan and Its Implication Get access, cit.

[184] Ibidem at 148 (explaining that Article 30-4 of Japan's Copyright Act allows unrestricted TDM uses as long as they involve extraction, comparison, classification, or other statistical analysis).

[185] Ibidem.

[186] See e.g. Japanese Agency for Cultural Affairs, General Understanding on AI and Copyright, 15 March 2024, pp. 3–5, available at: https://www.bunka.go.jp/english/policy/copyright/pdf/94055801_01.pdf (acknowledging that while Japan's Copyright Act allows for broad, non-consumptive uses—including AI training—the application of this exception is not absolute).

[187] See Matthew Sag, Fairness and Fair Use in Generative AI Authors, 92 Fordham L. Rev. 1887, 1917 (2024).

[188] E.g. it has been argued that this approach not only simplifies compliance for developers of generative AI but also resonates with the European notion of *Freier Werkgenuss* - a German doctrinal concept that excludes purely informational or non-expressive uses of a work from copyright protection. See Artha Dermawan, Text and Data Mining Exceptions in the Development of Generative AI Models: What the EU Member States Could Learn from the Japanese "Nonenjoyment" Purposes, 27 J. World Intell. Prop. 44, 54–56 (2023).

[189] See Regulation 3 of the Copyright and Rights in Performances (Research, Education, Libraries and Archives) Regulations 2014, No. 1372, which inserted Section 29A into the Copyright, Designs and Patents Act 1988. The Regulations entered into force on 1 June 2014. The scope of the exception was intended to align with the research exception under Article 5(3)(a) of the InfoSoc Directive.

[190] The so called "Hargreaves Review" was commissioned in December 2010 by the UK Prime Minister, Rt Hon. David Cameron MP, with the aim of developing proposals on how the UK's intellectual property framework could better support

must have already paid for a subscription or rely on openly licensed content. Moreover, the exception is explicitly non-commercial and does not extend to private-sector actors or commercial research.[191] Another significant limitation is the restriction on sharing: copies made under the exception cannot be transferred or communicated to other persons without the right holder's authorization. This impedes cross-institutional collaboration and undermines the scalability of research involving large, collaboratively mined corpora. The UK model includes an explicit prohibition on contractual override, which distinguishes it from many other jurisdictions. Section 29A(5) renders unenforceable any term of a contract that seeks to restrict acts permitted under the statutory exception.[192] While this does not solve the problem of technological protection measures (TPMs)—which UK law does not permit users to circumvent—it does protect the legal certainty of researchers against overreaching licensing practices. It is worth noting that, as of today, Section 29A of the CDPA remains unapplied.[193]

In addition, the United Kingdom has recently proposed a reform that mirrors the structure of Article 4 of the EU's CDSM Directive. Specifically, the 2024–2025 UK Government consultation proposes a commercial TDM exception, allowing copyright-protected content to be used for AI training unless rightsholders explicitly opt out.[194] This proposal - which de facto mimics the current EU approach - though presented as a step toward regulatory clarity and innovation, has been met with substantial scholarly criticism for effectively aligning copyright policy disproportionately with the interests of the AI industry, while neglecting its normative foundations and systemic coherence.[195] In particular, several common and pointed critiques have been raised. First, the opt-out model imposes a disproportionate and often unmanageable burden on rightsholders—particularly individual creators, educators, and public interest institutions—who may lack the technical or legal means to enforce their preferences effectively.[196] Second, the proposed exception reinforces existing asymmetries between technology companies and creators, enabling powerful actors to capture economic value from copyrighted works

entrepreneurialism, economic growth, and both social and commercial innovation. The final report, Digital Opportunity: A Review of Intellectual Property and Growth, was published on 18 May 2011. Available at https://assets.publishing.service.gov.uk/media/5a796832ed915d07d35b53cd/ipreview-finalreport.pdf

191 See e.g. Andres Guadamuz, A Scanner Darkly: Copyright Liability and Exceptions in Artificial Intelligence Inputs and Outputs, 73 GRUR Int. 111, 115–117 (2024) (discussing the narrow scope of the UK's TDM exception and its inapplicability to commercial generative AI training, which may lead to enforcement uncertainty and potential rights laundering risks).

192 See Copyright, Designs and Patents Act 1988, § 29A(5) (UK) ("To the extent that a term of a contract purports to prevent or restrict the making of a copy which, by virtue of this section, would not infringe copyright, that term is unenforceable").

193 See Eleonora Rosati, No step-free copyright exceptions: the role of the three-step in defining permitted uses of protected content (including TDM for AI-training purposes), 46 European Intellectual Property Review 262-274, 270 (2024).

194 See Secretary of State for Science, Innovation and Technology, UK Government Copyright and Artificial Intelligence Consultation. Available at https://www.gov.uk/government/consultations/copyright-and-artificial-intelligence

195 See e.g. Martin Kretschmer et al., Copyright and AI: Response by the CREATe Centre to the UK Government's Consultation (February 25, 2025). Available at SSRN: https://ssrn.com/abstract=5166928; Ann Kristin Glenster, et al., Written Evidence: UK Government Consultation on Copyright and Artificial Intelligence, Minderoo Centre for Technology and Democracy (2025). Available at https://www.repository.cam.ac.uk/handle/1810/381019; Gaetano Dimita et al., Response to the Copyright and AI Consultation (February 28, 2025). Queen Mary Law Research Paper No. 443/2025. Available at SSRN: https://ssrn.com/abstract=5164699; Zoi Krokida et al., Response to the public consultation of the UKIPO on Copyright and Artificial Intelligence (February 25, 2025). Available at SSRN: https://ssrn.com/abstract=5153968

196 See Martin Kretschmer et al., Copyright and AI: Response by the CREATe Centre to the UK Government's Consultation, cit.; Gaetano Dimita et al., Response to the Copyright and AI Consultation, cit.

without fair compensation or meaningful accountability.[197] Third, scholars warn that the UK approach treats copyright as a transactional obstacle to be streamlined, rather than as a framework of rights designed to protect expressive autonomy, cultural diversity, and economic justice.[198] It has been also noted that the UK Government's proposal risks accelerating a deregulatory shift where AI-generated content proliferates at the expense of human creativity, with minimal transparency, oversight, or remuneration structures.[199] Similarly, other commentators question why, in the face of demonstrable limitations within the EU's opt-out-based system, the UK would choose to replicate rather than rethink that approach.[200] Instead of crafting a context-sensitive regime aligned with the UK's own cultural, economic, and institutional traditions, the proposal appears to default to a flawed model already under strain across the Channel.[201] This alignment with existing, and arguably problematic, models may represent a missed opportunity for the UK to demonstrate regulatory leadership through the development of frameworks more closely attuned to its legal and economic specificities. As of now, the UK government has not made a final decision regarding the implementation of the proposed TDM exception. The outcome will depend on the government's assessment of the consultation feedback and its efforts to balance the interests of rightsholders and AI developers. Parliamentary debate on the matter remains contentious.[202] The UK's approach may offer a cautionary insight for the EU: without a renewed engagement with the underlying purpose and normative coherence of copyright—beyond the binary of access versus restriction—future reforms risk enabling extractive dynamics that could undermine both cultural ecosystems and public trust. In this light, comparative assessment should consider not only legal compatibility, but also the normative trajectory of copyright policy in the algorithmic age.

In stark contrast, the United States has never codified a TDM-specific exception. Instead, it relies on the more flexible doctrine of fair use under Section 107 of the Copyright Act of 1976. Landmark judicial

[197] See Martin Kretschmer et al., Copyright and AI: Response by the CREATe Centre to the UK Government's Consultation, cit.; Gaetano Dimita et al., Response to the Copyright and AI Consultation, cit.; Zoi Krokida et al., Response to the public consultation of the UKIPO on Copyright and Artificial Intelligence, cit.

[198] See Martin Kretschmer et al., Copyright and AI: Response by the CREATe Centre to the UK Government's Consultation, cit.; Ann Kristin Glenster, et al., Written Evidence: UK Government Consultation on Copyright and Artificial Intelligence, cit.; Gaetano Dimita et al., Response to the Copyright and AI Consultation, cit.; Zoi Krokida et al., Response to the public consultation of the UKIPO on Copyright and Artificial Intelligence, cit.

[199] See Gaetano Dimita et al., Response to the Copyright and AI Consultation, cit. at 1-2 (arguing that "AI-powered content generation will continue to sideline human creators, reinforcing existing power imbalances" and that innovation should not "come at the cost of human creativity").

[200] See European Writers' Council, Response to UK consultations: copyright and artificial intelligence (23 February 2025). Available at https://europeanwriterscouncil.eu/wp-content/uploads/2025/02/EWC-Submission-Copy-of-ANON-2HEH-VSN4-J.pdf

[201] See generally Zoi Krokida et al., Response to the public consultation of the UKIPO on Copyright and Artificial Intelligence, cit.; Martin Kretschmer et al., Copyright and AI: Response by the CREATe Centre to the UK Government's Consultation, cit.

[202] See Dan Milmo and Raphael Boyd, House of Lords pushes back against government's AI plans, The Guardian, 12 May 2025, available at https://www.theguardian.com/technology/2025/may/12/house-of-lords-pushes-back-ai-plans-data-bill

decisions, particularly Authors Guild v. Google[203] and Authors Guild v. HathiTrust,[204] have established that certain uses characteristic of Text and Data Mining (TDM)—such as full-text indexing, searchability, and analytical data extraction—can qualify as transformative under the U.S. fair use doctrine.[205] In both cases, courts held that digital reproductions of entire works—whether by Google or a consortium of university libraries—were justified because they enabled new insights, served public interest goals, and did not substitute for the original works in the market. The courts embraced the notion of "non-expressive use," affirming that using a copyrighted work solely to extract facts, patterns, or metadata does not infringe the exclusive rights of reproduction or communication. Moreover, these decisions underscore that commercial use does not preclude fair use, provided the new use is **transformative** and does not adversely impact the market for the original work. Notably, U.S. law does not restrict fair use to certain classes of users, nor does it impose constraints on sharing or preserving TDM corpora. However, practical obstacles remain. Contractual restrictions in licensing agreements may still limit access to digital content for TDM purposes, and circumvention of TPMs is prohibited under Section 1201 of the DMCA.[206] In order to mitigate the barrier posed by the DMCA's anti-circumvention rules, the U.S. Copyright Office adopted a narrow exemption in 2021 that allows researchers at non-profit educational or research institutions to bypass technological-protection measures (TPMs) solely to conduct text-and-data mining (TDM) for scholarly, non-commercial purposes.[207] More recently (May 2025), the Copyright Office issued the first government report on generative-AI training; it treats ingestion of copyrighted works as prima facie infringement and notes that fair-use outcomes remain uncertain, urging Congress to consider voluntary or statutory licensing solutions.[208] In the meantime, the U.S. framework appears to be evolving toward a market-based solution. Copyright law is, in fact, fundamentally based on an *opt-in* structure—that is, any use of protected content requires prior authorisation from the rightsholder, unless a clearly defined exception, limitation, or the fair use doctrine applies. While many high-profile legal disputes in the U.S. remain pending, several have already been resolved through settlement agreements—and more are likely to follow—indicating that licensing negotiations may emerge as the dominant path forward. This trend reinforces the notion that voluntary agreements, rather than categorical exceptions, are *de facto* shaping the operational landscape. However, this market-led model raises concerns about power asymmetries: large technology companies possess significant bargaining leverage, whereas smaller creators often lack the resources to effectively assert or enforce their rights in such negotiations.

---

[203] F.3d 202 (2d Cir. 2015).

[204] 755 F.3d 87 (2d Cir. 2014).

[205] Under U.S. law, a use is considered "transformative" if it adds something new, with a further purpose or different character, altering the original with new expression, meaning, or message. Courts have found that TDM qualifies as transformative where it does not reproduce expressive content for the same purpose, but instead extracts factual or structural information to serve a distinct analytical or informational goal. This distinction significantly weighs in favor of fair use under the first factor of the statutory test (17 U.S.C. § 107).

[206] Digital Millennium Copyright Act, Pub. L. No. 105-304, 112 Stat. 2860 (1998) (codified at 17 U.S.C. §§ 1201-1202 (2012).

[207] Exemption to Prohibition on Circumvention of Copyright Protection Systems for Access Control Technologies, Final Rule, 86 Fed. Reg. 59627, 59643 – 59645 (Oct. 28 2021) (codified at 37 C.F.R. § 201.40(b)(13)).

[208] See U.S. Copyright Office, Copyright and Artificial Intelligence, Part 3: Generative AI Training, cit.

Taken together, these international models point to a series of regulatory design choices that can inform EU policy. Japan's statutory exception provides a comprehensive model in which the law decouples the permissibility of TDM from both the user's identity and the purpose of use, framing data mining as a functional, non-expressive activity. [209] While its broad scope offers legal clarity, the model also raises questions regarding downstream uses and the treatment of licensing restrictions. The UK framework, although shaped by earlier EU-derived limitations, offers an instructive example in one respect: its explicit prohibition of contractual override, a dimension still insufficiently addressed under Article 4 of the EU's CDSM Directive. The U.S. experience, meanwhile, reflects a more fluid and market-driven approach. The fair use doctrine—developed judicially rather than legislatively—has shown adaptability in accommodating new technological uses, but it remains an open question whether such flexibility will persist in addressing the legal challenges posed by generative AI training. At the same time, the increasing reliance on settlement agreements and licensing negotiations suggests that, in practice, voluntary agreements are becoming a key mechanism for managing rights in the AI training context. This evolving situation offers valuable insights for jurisdictions considering how to balance legal certainty, user rights, and creative sector sustainability.

These dynamics are particularly visible in recent litigation concerning generative AI training in the UK. In the context of Getty Images v. Stability AI,[210] it has been noted that, although UK copyright law includes a non-commercial TDM exception under Section 29A of the CDPA, this exception does not extend to generative AI training by private companies.[211] As persuasively argued, requiring licences for such uses ensures that creators are fairly compensated, prevents freeriding, and may even foster collaborative innovation by promoting transparent licensing frameworks.[212] While the UK is no longer bound by EU copyright directives, the underlying concern—that unlicensed AI training undermines incentives for creation—resonates across jurisdictions and could inform ongoing discussions within the EU framework.

---

[209] See Artha Dermawan, Text and Data Mining Exceptions in the Development of Generative AI Models: What the EU Member States Could Learn from the Japanese "Nonenjoyment" Purposes, 27 J. World Intell. Prop. 44 (2023); Tatsuhiro Ueno, The Flexible Copyright Exception for 'Non-Enjoyment' Purposes - Recent Amendment in Japan and Its Implication Get access, 70 GRUR International 145-152 (2021).

[210] See Getty Images (US), Inc. v. Stability AI Ltd., [2023] EWHC (Ch) 3090 (UK High Court).

[211] See e.g. Zoya Yasmine, Getty Images v Stability AI: Why Should UK Copyright Law Require Licences for Text and Data Mining Used to Train Commercial Generative AI Systems, 1 Cambridge Journal of Artificial Intelligence 108-120 (2024); Paula Westenberger & Despoina Farmaki, Artificial Intelligence for Cultural Heritage Research: The Challenges in UK Copyright Law and Policy (Feb. 23, 2025), available at https://ssrn.com/abstract=5153757 (arguing that the current UK TDM exception is not fit for the purpose of AI training, particularly in real-world or public-private collaborative contexts). See also Secretary of State for Science, Innovation and Technology (2024) Consultation Outcome, A pro-innovation approach to AI regulation (CP 1019) Presented to Parliament by the Secretary of State for Science, Innovation and Technology by Command of His Majesty on 6 February 2024 https://assets.publishing.service.gov.uk/media/65c1e399c43191000d1a45f4/a-pro-innovation-approach-to-ai-regulation-amended-governement-response-web-ready.pdf (mentioning significant opposition from the creative industry to the UK Government proposal to adopt a EU style "opt out" copyright exception arguing that it undermines existing licensing frameworks and fails to ensure fair compensation for creators)

[212] See Zoya Yasmine, Getty Images v Stability AI: Why Should UK Copyright Law Require Licences for Text and Data Mining Used to Train Commercial Generative AI Systems, cit.

A further instructive example can be drawn from Switzerland, where the copyright exemption for research purposes is notably broader. As Picht and Thouvenin observe, Swiss law permits text and data mining for both scientific and commercial research and includes technical reproductions necessary for AI training, without the narrow limitations found in EU law.[213] This model offers another pragmatic alternative that balances innovation incentives with legal clarity—especially in cross-sectoral AI development contexts. These comparative observations are also consistent with the recent findings of Sag and Yu, who identify an emerging international equilibrium around non-expressive uses of copyrighted works for AI training purposes.[214] Their cross-jurisdictional survey suggests that many countries are converging—albeit unevenly—toward a middle-ground position that recognises the social utility of TDM and AI training without categorically permitting or banning unlicensed uses.[215] Importantly, the authors highlight three converging forces: the centrality of the idea–expression dichotomy,[216] global AI competition, and a regulatory "race to the middle."[217] Their work strengthens the argument that EU law should not simply tighten enforcement or expand exceptions in isolation, but rather adopt a granular, future-proofed framework that aligns with the technological character of generative AI and supports legally secure cross-border data practices. In parallel to these national and regional approaches, the World Intellectual Property Organization (WIPO) has, since 2019, convened a dedicated conversation on the implications of artificial intelligence for intellectual property.[218] This ongoing initiative has gathered governments, experts, and stakeholders from across the globe to examine pressing questions related to authorship, ownership, transparency, and liability. Its evolution reflects a growing recognition that generative AI challenges foundational IP concepts and demands regulatory innovation beyond existing territorial frameworks. As such, it underscores the importance of aligning EU reforms not only with internal market goals but also with emerging international principles and soft law standards.

Considering the diversity of regulatory models explored above, the EU's current TDM regime appears both fragmented and insufficiently future-proof. Article 4 of the CDSM Directive permits rightsholders to reserve their rights via machine-readable opt-outs, thereby undermining the effectiveness of the exception. While Article 7(1) prohibits contractual override, technological override remains permissible, as the Directive fails to provide a meaningful mechanism for researchers to challenge or circumvent TPMs that block otherwise lawful TDM activities. Moreover, the EU's failure to clearly permit

[213] See Peter Georg Picht and Florent Thouvenin, AI and IP: Theory to Policy and Back Again – Policy and Research Recommendations at the Intersection of Artificial Intelligence and Intellectual Property, 54 IIC 916, 928 (2023).

[214] See Matthew Sag and Peter K. Yu, The Globalization of Copyright Exceptions for AI Training, 74 Emory Law Journal, 1–58 (2025).

[215] Ibidem.

[216] The idea-expression dichotomy is a fundamental principle in copyright law whereby protection applies only to the specific expression of an idea, not to the idea itself. While not expressly mentioned, this distinction is implicit in the Berne Convention, which protects "literary and artistic works" as expressions, but does not extend to ideas, procedures, or concepts. The dichotomy was developed doctrinally and jurisprudentially, notably in the U.S. case Baker v. Selden, 101 U.S. 99 (1879).

[217] See Matthew Sag and Peter K. Yu, The Globalization of Copyright Exceptions for AI Training, cit.

[218] See WIPO, WIPO Conversation on Intellectual Property and Frontier Technologies, available at: https://www.wipo.int/en/web/frontier-technologies/frontier_conversation

commercial TDM under conditions of legal certainty risks stifling research-driven innovation in the private sector and may generate a chilling effect for AI developers operating within the Union.

2.2. Implementation across Member States

Recent comparative findings published by the Communia Association (2024),[219] along with supplementary legal analyses and national reports, reveal a highly fragmented and uneven implementation of Articles 3 and 4 of the CDSM Directive across EU Member States.[220] This legal fragmentation underscores that the boundaries of what qualifies as lawful TDM—particularly in the context of commercial AI training—remain unsettled. For example, Germany's transposition and judicial interpretation suggest a more permissive stance in the absence of valid opt-outs, illustrating the lack of harmonised application across jurisdictions. This patchwork of national approaches introduces substantial variability in the interpretation and operationalisation of text and data mining (TDM) exceptions, with far-reaching consequences for both the scientific research ecosystem and the rapidly evolving field of generative AI. Although Articles 3 and 4 were intended to harmonise core aspects of TDM, particularly through the creation of mandatory exceptions for research and general-purpose data processing, the current implementation landscape reveals stark inconsistencies in scope, conditions, enforcement, and technical interoperability.

A significant number of Member States—nineteen, according to the Communia study—have opted to preserve or expand the broader research exceptions under Article 5(3)(a) of the InfoSoc Directive. Among them, eight countries (such as Croatia, Estonia, Latvia, and Slovakia) have introduced open-ended provisions that accommodate a wide range of TDM-related acts carried out by any user for non-commercial scientific purposes. These frameworks tend to offer more permissive legal environments, better suited to supporting open research collaborations and data-intensive analytical methods, including AI development. Notably, five Member States (including Germany and Hungary) explicitly allow the public dissemination of TDM outputs, either directly under Article 3 or through other applicable research exceptions. This reflects a pragmatic response to the growing need for transparency, reproducibility, and open sharing of datasets in AI training pipelines, and contrasts with the more restrictive formulation of Article 3 at the EU level.

Divergence becomes even more pronounced when examining the implementation of Article 4, which governs TDM for all purposes, including commercial AI development. The provision's opt-out mechanism, which permits rightsholders to reserve their rights through "machine-readable means," has been transposed in markedly different ways. For instance, while eleven Member States (such as Belgium, Czech Republic, Latvia, and Slovenia) have mandated the use of technical protocols—like metadata tags or robots.txt files—other countries, including France and Italy, have transposed the opt-

[219] See Teresa Nobre, The Post-DSM Copyright Report: research rights, February 5, 2024. Available at https://communia-association.org/2024/02/05/the-post-dsm-copyright-report-research-rights/

[220] See e.g. Study for European Commission: Directorate-General for Research and Innovation – Improving Access to and Reuse of Research Results, Publications and Data for Scientific Purposes, Brussels: Publications Office of the European Union 2024, available at: https://data.europa.eu/doi/10.2777/633395, (noting that Member States have implemented TDM exceptions inconsistently creating legal uncertainty for researchers and developers across Europe)

out requirement with little to no guidance on technical implementation. In France, rights holders have relied on contractual terms of service: [221] in practice, rights holders like SACEM have exercised opt-outs by including reservations in their terms of service. However, there is no standardized, machine-readable format mandated or widely adopted, leading to ambiguity about the effectiveness and enforceability of such opt-outs. Italy's implementation closely mirrors the text of Article 4 of the CDSM Directive but does not provide specific guidance on how rights holders should express opt-outs. There is no mention of machine-readable formats or standardized procedures, resulting in uncertainty for both rights holders and TDM users regarding the validity and recognition of opt-out declarations.[222] Spain similarly implemented Article 4 via Article 67 of Royal Decree-Law 24/2021, but its approach has been criticized for failing to clearly extend the TDM exception to all relevant neighbouring rights, and for lacking explicit rules on how rights holders should express opt-outs.[223] The absence of technical detail and sectoral guidance has raised concerns about legal certainty, particularly in the context of large-scale AI training.[224] In contrast, Germany has introduced a clear obligation for opt-outs to be machine-readable,[225] as confirmed by the Hamburg District Court in the LAION case,[226] though the court controversially accepted natural-language disclaimers as potentially compliant, illustrating the interpretive fluidity even within relatively structured regimes. The Netherlands also provides a notable example: [227] its courts have upheld the validity of TDM under Article 4 where no machine-readable opt-out was implemented, reaffirming that legal clarity hinges on strict technical compliance.[228] All these disparities create a compliance minefield for AI developers, particularly those using automated tools to ingest large-scale web data, who must assess the legal status of each source on a jurisdiction-by-jurisdiction basis.

Compounding the complexity are differing national positions on enforcement and technological protection measures (TPMs). While EU law provides robust protection for TPMs under Article 6 of the InfoSoc Directive, only a handful of Member States have adopted corresponding safeguards to ensure that lawful uses under copyright exceptions are not obstructed by digital locks. Slovenia stands out

---

[221] France transposed Article 4(3) into Article L122-5-3 of its Intellectual Property Code.

[222] Italy transposed Directive (EU) 2019/790 through Legislative Decree No. 177 of Nov. 8, 2021, which introduced Article 70-quater into Law No. 633 of Apr. 22, 1941, Legge sul diritto d'autore (Italian Copyright Act), thereby implementing the general text and data mining exception into Italian law.

[223] See Article 67, Royal Decree-Law 24/2021 (Spain), amending the Spanish Copyright Act.

[224] Teresa Nobre, A First Look at the Spanish Proposal to Introduce ECL for AI Training, Kluwer Copyright Blog (Dec. 11, 2024), https://copyrightblog.kluweriplaw.com/2024/12/11/a-first-look-at-the-spanish-proposal-to-introduce-ecl-for-ai-training/ (criticizing the overlap between Spain's proposed ECL scheme and the existing TDM exception under Article 4, and noting shortcomings in Spain's implementation of both Articles 3 and 4 of the DSM Directive).

[225] Germany's implementation of Article 4 of the CDSM Directive is codified in Section 44b of the Urheberrechtsgesetz (UrhG).

[226] District Court of Hamburg, Robert Kneschke v. LAION e.V., Case No. 310 O 227/23.

[227] The Netherlands transposed Article 4 of Directive (EU) 2019/790—governing the general text and data mining (TDM) exception—into its national law through Article 15o of the Dutch Copyright Act (Auteurswet).

[228] See Amsterdam District Court, DPG Media et al. v. HowardsHome, ECLI:NL:RBAMS:2024:6563 (Nov. 15, 2024) (holding that TDM use was lawful under Article 15o of the Auteurswet in the absence of a machine-readable opt-out); see also "Dutch Court Holds That TDM Opt-Out Must Be Done by 'Machine-Readable' Means," The IPKat (Feb. 2025), https://ipkitten.blogspot.com/2025/02/dutch-court-holds-that-tdm-opt-out-must.html

with a notably progressive provision: it requires rights holders to disable TPMs within 72 hours of receiving a legitimate request to allow lawful TDM activities.[229] This enforcement mechanism gives practical effect to user rights and reduces the friction between copyright exceptions and access control technologies. In contrast, most Member States provide no such obligation or enforcement path, leaving the rights of lawful users—such as researchers or AI developers—largely theoretical when faced with locked digital content.

Further insight into the causes and consequences of this fragmentation can be found in the broader legal literature. According to a recent comparative study on the implementation methodology of the DSM Directive, the flexibility granted to Member States in transposing exceptions has led to both literal transpositions and broader "gold-plating" practices.[230] The concept of "lawful access," a cornerstone of both Articles 3 and 4, has not been uniformly interpreted.[231] Some Member States, like Slovenia and Poland, have adopted restrictive definitions. For instance, Slovenia's legislation excludes freely accessible online content from the scope of lawful access[232]—despite the guidance of Recital 14—while Poland prohibits any TDM use with a commercial purpose under the research exception and introduces ambiguous language that could exclude common pre-processing steps from protection.[233] These interpretations stand in tension with both the spirit and the text of the Directive and may risk incompatibility with EU law.

The cumulative effect of these disparities is a troubling degree of legal uncertainty for researchers and developers engaged in TDM. Activities that are fully lawful in one Member State may constitute infringement in another, depending on how national legislatures have implemented and interpreted key provisions regarding opt-outs, enforcement rights, and the definition of lawful access. This undermines the fundamental goals of the Digital Single Market, particularly the principle of cross-border portability for research and innovation, and calls into question the EU's strategic objective of leading the world in the development of trustworthy, rights-compliant AI. Without renewed harmonisation efforts—particularly in the areas of opt-out standardisation, lawful access definitions, and enforceable rights to circumvent obstructive TPMs—EU copyright law risks becoming not an enabler of technological advancement, but a structural obstacle to it. This concern has already been acknowledged by the

[229] See Maja Bogataj Jančič, Exceptions with teeth: the new Slovenian text and data mining provisions, knowledgerights21 (October 5, 2023). Available at https://www.knowledgerights21.org/news-story/exceptions-with-teeth-the-new-slovenian-text-and-data-mining-provisions/

[230] See Branka Marušić, TDM Exception or Limitation –Methodology of Implementation in the EU Member States: Creating Cohesion or Diversion?, Stockholm IP Law Review 2024#2, 19-24 (April 2025).

[231] See e.g. Matthew Sag, Fairness and Fair Use in Generative AI Authors, 92 Fordham L. Rev. 1887, 1917/18 (2024) (Although writing in the context of U.S. fair use doctrine, Sag notes that lawful access remains a distinct prerequisite, not automatically satisfied even where subsequent use is non-expressive or transformative. He cautions against elevating lawful access to a per se requirement, particularly when access through legal markets is unavailable or conditioned on restrictive licensing—raising questions that resonate across jurisdictions in the context of AI training and text/data mining).

[232] See Maja Bogataj Jančič and Ema Purkart, Text and Data Mining in the Slovenian Legal System, Stockholm IP Law Review 2024#2, 5-8 (April 2025).

[233] See Konrad Gliściński, Polish Implementation of TDM Exceptions– General Characteristics, Stockholm IP Law Review 2024#2, 9-18 (April 2025).

European Parliament in its 2020 resolution, which stressed the importance of establishing a harmonised EU regulatory framework for AI and intellectual property, preferably in the form of a regulation to avoid fragmentation across Member States.[234]

2.3. Impact on rightsholders

The use of copyright-protected works to train generative AI systems has led to widespread concerns among authors, performers, and other rightsholders regarding the lack of consent, attribution, and above all, remuneration. While Article 4 of the Directive on Copyright in the Digital Single Market (CDSM Directive) permits text and data mining (TDM) by default—unless rightsholders opt out via machine-readable means—this exception fails to provide any form of compensation. As a result, a structural "value gap" has emerged between the commercial benefits accrued by AI developers and the lack of financial return for the human creators whose works underpin these systems.

Creators' groups argue that the current framework allows generative AI developers to benefit from mass-scale ingestion of creative works without returning any value to the original contributors.[235] In 2023, a broad coalition of European authors and performers urged EU lawmakers to include safeguards in the AI Act to ensure that generative AI technologies do not displace or devalue human creativity without compensation.[236] Their demands encompass not only consent and transparency, but also enforceable remuneration rights, either through collective licensing schemes or new statutory mechanisms.

In response, several proposals have been advanced to bridge this gap. One is the establishment of collective management organisations that could offer blanket licences for AI training purposes, distributing fees among a broad base of rightsholders.[237] This model draws on well-established practices in music and broadcasting and could provide a scalable solution for dataset licensing. A more ambitious proposal involves the introduction of a new EU-level right to remuneration for authors whose works are used in training AI systems, analogous to the press publishers' right under Article 15 CDSM.[238] Such a right could ensure income flows even where direct licensing is impractical.

---

[234] See European Parliament resolution of 20 October 2020 on intellectual property rights for the development of artificial intelligence technologies, 2020/2015(INI), 2021 O.J. (C 404) 129, at §3.

[235] See supra note 148.

[236] See Initiative Urheberrecht, Authors and Performers Call for Safeguards Around Generative AI in the European AI Act, 19 April 2023, available at https://urheber.info/media/pages/diskurs/call-for-safeguards-around-generative-ai/c93a5ab197-1681904353/final-version_authors-and-performers-call-for-safeguards-around-generative-ai_19.4.2023_12-50.pdf

[237] See Martin Senftleben, Generative AI and Author Remuneration, 54 IIC – International Review of Intellectual Property and Competition Law 1535 (2023) (proposing a levy-based remuneration scheme administered by collective management organizations, focusing on the output of generative AI systems and aiming to compensate authors for market substitution effects while supporting human creativity).

[238] See Christophe Geiger and Vincenzo Iaia, The forgotten creator: Towards a statutory remuneration right for machine learning of generative AI, 52 Computer Law & Security Review 1–9 (2024) (proposing a statutory license model grounded in fundamental rights to ensure fair remuneration for authors whose works are used to train generative AI, balancing innovation incentives with the protection of creators' material and moral interests).

However, both proposals face significant feasibility challenges. Collective licensing would require large-scale rights aggregation and coordination across sectors and Member States—something that is currently lacking for literary, visual, or multimedia works. Moreover, defining the scope and pricing of such blanket licences for AI training (a use unlike traditional consumption) presents novel legal and economic difficulties. Similarly, a new remuneration right for AI training would likely require EU-level legislation, raising questions about its compatibility with existing copyright architecture, its enforceability across jurisdictions, and the risk of unintended consequences (e.g. overreach, excessive burdens on smaller developers).

Tech companies and innovation advocates have pushed back on these proposals, warning that imposing licensing or remuneration obligations for every work ingested into an AI training dataset could make AI development prohibitively expensive. They argue that such a regime risk creating gatekeeping power for large rightsholders, chilling innovation and entrenching incumbents. They also invoke the analogy of human learning, suggesting that AI systems "read" and "learn" from texts and images in ways that should be considered non-consumptive and therefore exempt from compensation.[239]

Market-based alternatives—such as voluntary licensing agreements—have begun to emerge. For example, Shutterstock has entered into a content licensing arrangement with OpenAI for its DALL-E image generation system, offering contributor compensation.[240] These voluntary models demonstrate that remuneration is technically feasible and can align incentives, but they remain limited in scope and unlikely to scale without regulatory intervention. Not all developers engage in such practices, and high-value datasets remain largely unlicensed. Additionally, recent legal scholarship has argued that offering generative AI systems to EU users—particularly where the output can reproduce parts of the training data—may constitute a "making available to the public" under Article 3(1) of the InfoSoc Directive. This perspective strengthens the enforcement potential of EU copyright law, even in cases where the training occurs outside the Union's territory.[241]

At the same time, rightsholders have begun defending their rights in court, including within the EU. For example, in November 2024 and January 2025, GEMA—Germany's largest collective management organisation—initiated legal proceedings against OpenAI and Suno, alleging that their generative AI

[239] While this analogy is frequently invoked to frame AI training as a non-consumptive, human-like learning process, this study maintains that such a comparison is generally untenable under EU copyright law. The ingestion of protected works by generative AI systems typically involves acts of reproduction that extend beyond analytical use (see Section 2.1.2). Moreover, from a doctrinal standpoint, AI systems lack the cognitive features that justify exceptions for human learning: unlike human authors, who understand and reinterpret ideas within a conceptual framework, AI systems operate *agere sine intelligere*—they act without understanding (See Luciano Floridi, AI as Agency Without Intelligence: On ChatGPT, Large Language Models, and Other Generative Models, cit). This cognitive gap has profound legal implications. Human learners can restate an idea in a novel way without infringing copyright, thanks to the idea/expression dichotomy. In contrast, AI systems must ingest, copy, and statistically process the actual expressions of works in order to generate outputs. As such, even where no recognisable similarity exists between the training data and the output, this does not alter the legal characterisation of the training process itself as involving protected acts of reproduction.

[240] See Shutterstock Expands Partnership with OpenAI, Signs New Six-Year Agreement to Provide High-Quality Training Data, Press release (July 11, 2023). Available at https://investor.shutterstock.com/news-releases/news-release-details/shutterstock-expands-partnership-openai-signs-new-six-year

[241] On this, see Tim Dornis, Generative AI, Reproductions Inside the Model, and the Making Available to the Public, cit.

systems unlawfully retained and exploited protected works during the training process.[242] These cases highlight the growing legal pushback from rightsholders and underscore the perceived insufficiency of existing safeguards under the TDM framework.

The final recommendations of this study therefore favour a more pragmatic, incremental approach (Sections 4.1 and 4.2). While the concerns of rightsholders are legitimate and urgent, the feasibility of implementing robust remuneration schemes across the EU remains uncertain. Structural reforms may be required, but must be accompanied by technical standards, stakeholder coordination, and phased legal development. Without such groundwork, there is a risk of designing legal mechanisms that are aspirational but unworkable in practice. As detailed in Recommendation 4.3, the study advises exploring compensation models adapted to AI training, such as voluntary collective licensing schemes and revenue-sharing mechanisms, while recognising the legal and practical barriers to implementing a statutory remuneration right at this stage. A future-proof solution must balance innovation with fairness, and realism with ambition.

2.4. Author's Rights and remuneration for AI training uses

While the CDSM Directive's Article 4 establishes a mandatory, fee-free exception for bona fide Text and Data Mining (TDM), the analysis in Sections 2.1.1 and 2.1.2 demonstrates that the training of generative AI models—far from merely extracting factual patterns or semantic insights—falls outside the very contours of what EU law envisages as TDM.[243] Generative systems do not simply extract statistical correlations from data; they process and model the expressive patterns embedded in copyrighted works in order to compute outputs—such as text or images—that may resemble human-authored content. Consequently, the common practice of invoking the Article 4 exception as a blanket legal basis for large-scale AI training rests on a misapplication of the TDM: the technical processes at play more closely resemble reproduction and transformative reuse than the knowledge-extraction activities the Directive was designed to facilitate.

Yet, despite the conceptual mismatch, many AI developers have proceeded as if generative training squarely fell within the TDM exception, effectively placing creative works into their models without authorisation, notification, or remuneration.[244] From the perspective of rightsholders, this disconnect amounts to a policy vacuum. The European Parliament already acknowledged this gap in its 2020 resolution, stressing the importance of fair remuneration for authors whose works are used in AI

[242] See Gema Press release, Fair remuneration demanded: GEMA files lawsuit against Suno Inc. (Jan., 21, 2025). Available at https://www.gema.de/en/w/press-release-lawsuit-against-suno; Gema Press release, GEMA files model action to clarify AI providers' remuneration obligations in Europe (Nov. 13, 2024). Available at https://www.gema.de/en/w/gema-files-lawsuit-against-openai

[243] See Christophe Geiger, et al., Text and Data Mining in the Proposed Copyright Reform: Making the EU Ready for an Age of Big Data?, 49 IIC 814, 818–819 (2018) (emphasizing that the EU TDM exception was designed to facilitate scientific and analytical innovation, not the large-scale appropriation of expressive content for commercial purposes).

[244] See OECD, Intellectual Property Issues in Artificial Intelligence Trained on Scraped Data, cit. at 14 (highlighting that data scraping frequently occurs without the consent of rights holders, raising risks of copyright infringement, database rights violations, and breaches of publicity and moral rights).

systems, and warning that the digital economy must not erode the foundational incentives for human creativity.[245]

Under the current framework, authors face a stark binary choice: they may deploy technological protection measures (TPMs) to "opt out" entirely—thus sterilizing their works from inclusion in any automated analysis—or passively allow unfettered use of their creations, with no right to be informed or compensated when their labour fuels multimillion-dollar AI products. There is no intermediary route by which a creator can expressly grant permission for AI training while negotiating fair payment or attribution. In economic terms, this legal vacuum is compounded by a profound asymmetry in bargaining power. Individual authors, particularly freelancers and small creators, have limited capacity to negotiate licensing terms or to monitor and enforce their rights, especially against large AI developers and platforms with vast technical and legal resources. This disparity creates a coercive dynamic: either accept unremunerated use of one's work for AI training, or risk cultural and economic irrelevance. Such conditions effectively deprive creators of meaningful agency, turning consent into a formality rather than a genuine choice. From a regulatory perspective, this imbalance constitutes a textbook market failure—one in which voluntary agreements are neither fair nor freely negotiated, and where rights are systematically under-enforced.[246] Any future remuneration framework must therefore not only address the absence of compensation, but also rebalance negotiating conditions to empower authors vis-à-vis platform operators and AI developers. This regulatory impasse is not only a legal shortcoming—it reflects deeper structural asymmetries in the creative economy that require targeted redress. Beyond the challenges of ex ante licensing, a second structural concern arises at the distribution level. As generative systems become capable of producing vast quantities of plausible, low-cost content, there is a growing risk that automated outputs will crowd out human authorship in digital marketplaces, streaming platforms, and algorithm-driven content feeds.[247] This saturation effect not only distorts discoverability and remuneration, but also threatens to relegate human creators to a residual role—serving merely as raw data providers for AI systems rather than autonomous contributors to public discourse and culture.[248] Without safeguards that ensure visibility, attribution, and market access for human-generated works, the promise of creative diversity risks being supplanted by the scale advantages of synthetic expression.

---

[245] See European Parliament resolution of 20 October 2020 on intellectual property rights for the development of artificial intelligence technologies, 2020/2015(INI), 2021 O.J. (C 404) 129, at Recital D and §6.

[246] See e.g. U.S. Copyright Office, Copyright and Artificial Intelligence, Part 3: Generative AI Training, cit. at 92; 103 (discussing barriers to meaningful compensation and the risks of market exclusion for individual creators). See also, Katherine Lee et al., Talkin' 'Bout AI Generation: Copyright and the Generative-AI Supply Chain, 71 Journal of the Copyright Society of the U.S.A. (forthcoming 2024), at. 35, 77–79 (analyzing the extraction of expressive works without compensation and the substitution risks posed by AI-generated outputs);

[247] U.S. Copyright Office, Copyright and Artificial Intelligence, Part 3: Generative AI Training, cit. at 64; 103 (discussing the risk that AI outputs may displace human works, erode licensing value, and concentrate exposure on automated content).

[248] Ibidem. See also Tim W. Dornis, The Training of Generative AI is Not Text and Data Mining, cit. at 65–66 and 70–71 (criticizing the unlicensed ingestion of copyrighted works for AI training, and discussing how generative systems are designed to replicate expressive content in a way that competes with human authorship); Pamela Samuelson, Generative AI Meets Copyright, cit. at 158–159 (noting creators' lack of compensation and control over AI training, and the risk that AI outputs will displace human-authored works in creative and licensing markets).

These dual pressures—first at the licensing level, and second at the distribution layer—underscore the necessity of rethinking the existing framework and its underlying assumptions. This all-or-nothing regime has provoked intense debate over fairness and the proper allocation of value in the digital age. If generative AI truly lies beyond the scope of TDM, then the very legal justification for uncompensated training evaporates—and yet, without a clear alternative legal basis authorising large-scale AI ingestion of copyrighted content, both innovators and authors find themselves operating in a legal grey zone. Against this backdrop, the question of authors' rights and remuneration becomes pressing: How can EU policy reconcile the need for access to vast, high-quality datasets that drive AI innovation, with the equally legitimate demand that creators share in the economic returns generated by the use of their works? Recent trends in the licensing market further illustrate this complexity. As highlighted in the recent 2025 EUIPO study, a growing number of agreements between GenAI developers and rightsholders—particularly in publishing, image, and music sectors—reflect shifting dynamics shaped by data quality, metadata richness, annotation costs, and the role of intermediaries.[249] These factors, alongside concerns over synthetic data and dataset substitution, influence how value is distributed and who benefits. Such developments suggest that any policy response must account not only for legal design but also for the evolving realities of data-driven market structures.

The following pages explore the contours of this debate, mapping stakeholders' positions and surveying potential mechanisms—ranging from voluntary licensing and collective bargaining to new statutory remuneration entitlements—that might restore balance without unduly stifling technological progress.

2.4.1. Regulatory Gaps and Remuneration Challenges

As discussed above, the current TDM framework offers no practical pathway for negotiated consent or remuneration. Article 4 of the CDSM Directive establishes a mandatory, fee-free exception that applies by default unless rightsholders actively opt out. Unlike other EU copyright exceptions—such as the private copying exception, which is paired with a levy to ensure compensation—this provision imposes no duty to inform, credit, or remunerate authors when their works are repurposed for automated analysis. In practice, this framework leaves creators without any enforceable mechanism to authorize, deny, or license the use of their works for AI training under negotiated terms.

Stakeholder reactions to this legal lacuna divide sharply along traditional fault lines. Creators' associations—from the European Writers' Council to federations of visual artists and musicians—denounce the uncompensated appropriation of their works as a modern "value gap."[250] They point out that generative AI platforms reap substantial commercial rewards by leveraging professional-grade content in their models, yet the originators of that content see nothing but the residual risk of

[249] See EUIPO, The Development of Generative Artificial Intelligence from a Copyright Perspective (May 2025), cit. at 90–95, 107–114 (noting the role of data quality, metadata, synthetic data, and platform intermediation in shaping licensing market dynamics and bottlenecks).

[250] See European Composer and Songwriter Alliance/European Writers' Council et al, Joint Statement from Authors' and Performers' Organizations on Artificial Intelligence and the AI Act (February 9, 2023), https://composeralliance.org/media/1136-joint-statement-on-ai-and-the-ai-act.pdf

displacement.[251] In late 2023, a coalition of leading European authors and performers submitted an open letter to the European Commission as part of the AI Act consultations, calling for built-in safeguards to ensure authors are neither deprived of income nor stripped of control when their works feed AI training.[252] Their proposals span from collective licensing schemes and revenue-sharing agreements to the creation of a bespoke remuneration right for TDM.[253] By contrast, the technology sector and innovation advocates argue that requiring ex ante licensing or micropayments for every individual work would render AI research logistically and economically unviable. They warn that a rights-clearance regime for training data would spawn prohibitive administrative costs and legal complexity, essentially granting legacy publishers and large cultural conglomerates gatekeeping power over the very inputs that drive new AI ventures. They frequently invoke the analogy of human learning—arguing that people absorb ideas, styles, and facts from reading and listening without owing micropayments to each author they learn from—and contend that automated model training ought to be regarded similarly as a "non-consumptive" use.[254]

Reconciling these positions demands inventive policy design. Several potential mechanisms have been proposed:

Voluntary Licensing and Content Partnerships. Private agreements between AI developers and content platforms can channel remuneration to creators without mandating state-imposed fees. The mid-2023 Shutterstock–OpenAI deal for supplying curated imagery to DALL·E illustrates how revenue-sharing models can emerge organically.[255] However, reliance on voluntary markets risks leaving less commercially visible works unlicensed and underserved.

Extended Collective Licensing. By empowering collecting societies to negotiate blanket TDM licenses on behalf of their memberships, Member States could replicate the radio and television music-licensing model. Such schemes would cover all works in a given repertoire—unless individual authors opt out—and distribute royalties according to usage. Crafting Extended Collective Licensing schemes for AI training would likely require legislative amendments to clarify societies' mandates and to establish equitable distribution keys.

Statutory Remuneration Right for TDM. Analogous to the press publishers' right under Article 15 CDSM, a new exclusive right could obligate AI practitioners to pay a levy or share of profits when copyrighted works are used in model training. While this approach promises comprehensive coverage,

---

[251] Ibidem

[252] Copyright Initiative, Authors and Performers Call for Safeguards Around Generative AI (April 20, 2023), https://urheber.info/media/pages/diskurs/call-for-safeguards-around-generative-ai/069a7d264a-1697140342/authors-and-performers-call-for-safeguards-around-generative-ai_20.4.2023.pdf

[253] See Martin Kretschmer, et al., Copyright Law and the Lifecycle of Machine Learning Models, 55 IIC – International Review of Intellectual Property and Competition Law 110 (2024) (acknowledging the potential of collective licensing to reduce market entry barriers, but highlighting significant challenges).

[254] See supra note 239.

[255] See Shutterstock Expands Partnership with OpenAI, Signs New Six-Year Agreement to Provide High-Quality Training Data, Press release (July 11, 2023). Available at https://investor.shutterstock.com/news-releases/news-release-details/shutterstock-expands-partnership-openai-signs-new-six-year

it raises complex questions about how to measure each work's "contribution" to an AI model and how to avoid duplicative payments when the same content appears in multiple datasets.

Moral Rights and Source Acknowledgment. Beyond financial compensation, authors may seek mechanisms for recognition and protection of their personal connection to the work. The current AI Act transparency requirement (Article 53(1)(c)—which mandates disclosure of the categories of copyrighted data used in high-risk AI systems—represents a first step. However, a more robust regime might grant authors the right to query providers directly or to receive automated notices when their works are ingested, thereby fostering accountability and potentially catalysing licensing discussions. In parallel, moral rights—particularly the right of integrity—are emerging as a distinct regulatory concern. A recent EU study found that even when AI outputs do not reproduce original content in a recognisable way, they may still infringe moral rights if they mimic the author's style or cause reputational harm.[256] This perspective was endorsed by 67% of surveyed experts, who supported allowing rightsholders to invoke moral rights to oppose AI training, even where economic rights exceptions like TDM might apply.[257] These findings suggest that policy responses should not be limited to transparency and remuneration mechanisms but also address normative safeguards for attribution, reputation, and personal dignity, especially in sensitive fields such as literature, political speech, and the visual arts.

Two detailed scholarly proposals have also emerged, offering alternative remuneration architectures—yet each faces considerable practical challenges when assessed in light of EU copyright law and the realities of AI deployment.

The first proposal, advanced by Geiger and Iaia,[258] envisions the **introduction of a statutory licence** specifically tailored to machine-learning purposes. Under this model, any commercial use of copyrighted works to train generative AI would automatically trigger a mandatory licence, thereby dissolving the need for individual permissions or the blanket opt-out mechanism of Article 4(3) CDSM. Remuneration rates would be calibrated either through collective bargaining by authors' societies or set ex ante by a dedicated regulator, applying the "appropriate and proportionate" criteria already established in the CDSM Directive. Collected fees would flow into social and cultural funds managed by collecting societies, ensuring that creators receive direct and ongoing support. This licence is grounded in fundamental-rights reasoning—balancing the public's right to science and culture against authors' moral and material interests—and would embed a digital-constitutional framework into EU copyright governance.

Despite its theoretical elegance, the statutory-licence approach confronts many obstacles. First, accurately valuing each work's contribution to an opaque, high-dimensional training corpus is practically impossible, risking arbitrary fee schedules and litigation over rate-setting. Second, AI developers tightly guard their training pipelines as trade secrets; a licence premised on full

[256] European Commission, Study on Copyright and New Technologies: Copyright Data Management and Artificial Intelligence, cit. at 230.

[257] Ibidem at 228-30.

[258] See Geiger Christophe and Iaia Vincenzo, The forgotten creator: towards a statutory remuneration right for machine learning of generative AI. 52 Computer Law & Security Review 1-9 (2024).

transparency of dataset composition would undermine commercial confidentiality and complicate web-scale crawling. Third, the very notion of a paid licence conflicts with the CDSM Directive's explicit design of the TDM exception as mandatory and fee-free—rectifying this would demand a wholesale legislative overhaul of the Directive itself. Finally, without uniform EU-level oversight, Member States might adopt disparate licence regimes, generating a fragmented legal landscape and deterring smaller innovators with prohibitive compliance costs.

In contrast, Senftleben's second proposal sidesteps input-level complexity by imposing an **output-oriented "AI levy"** on providers of generative systems whose outputs could substitute for human creations.[259] Drawing on analogies to the phonogram levy in the Rental and Lending Directive,[260] this approach would require any commercial AI service whose outputs reach a threshold of human-like substitutability to pay a lump-sum levy—calculated, for example, as a percentage of turnover, user subscriptions, or volume of generated content. The pooled funds would be distributed by collecting societies to support authors' livelihoods, finance training programmes, and underwrite new creative projects. By decoupling remuneration from specific training datasets, the levy avoids the secrecy concerns of the statutory licence and transforms AI-generated revenue into resources for human creators.

Yet the output levy, too, is fraught with implementation challenges. Defining and evidencing the "potential to substitute" human creativity is legally and technically indeterminate, rendering enforcement highly subjective. Setting a levy rate that both delivers meaningful support to authors and preserves the EU's attractiveness as an AI hub requires economic data that does not exist, risking either under-collection or economic deterrence. Moreover, administering a novel lump-sum mechanism would impose substantial new burdens on collecting societies, which must develop audit, collection, and repartitioning frameworks far beyond their current remit. Finally, by penalizing AI deployment in general rather than targeting specific uses, an output levy could unintentionally incentivize platform relocation to jurisdictions without such levies, undermining the EU's broader digital strategy.

While both the statutory-licence and AI-levy proposals offer principled routes to closing the 'value gap' between generative AI platforms and creative rightsholders, they each face substantial challenges in terms of legal coherence and practical feasibility.

A more technical and economically driven proposal envisions a token-based royalty system grounded in the marginal utility of training data. This approach uses influence measurements, such as Shapley-value approximations, to assess the contribution of individual content units (e.g., text tokens) to model performance, and proposes distributing royalties proportionally.[261] While not grounded in current legal practice, this model represents an innovative economic alternative that could complement legal

[259] See Martin Senftleben, Generative AI and Author Remuneration. 54 IIC – International Review of Intellectual Property and Competition Law, 1535 (2023).

[260] Directive 2006/115/EC of the European Parliament and of the Council of 12 December 2006 on rental right and lending right and on certain rights related to copyright in the field of intellectual property (codified version) (27 December 2006) (Rental and Lending Directive), OJ L 376.

[261] See Jiachen T. Wang et al., An Economic Solution to Copyright Challenges of Generative AI, arXiv (Apr. 2024), https://arxiv.org/abs/2404.13964

proposals by introducing granular and algorithmically computable methods of allocating value to rightsholders in AI training contexts.

Together, all these proposals underscore the need for a hybrid regulatory vision—one that combines the legal structure of collective rights management with the adaptive potential of data-driven allocation models. Any future policy must therefore blend elements of both approaches—perhaps through enhanced transparency requirements, targeted voluntary licensing pilots, and a light-touch collective framework—to ensure that Europe's copyright system evolves in tandem with its ambitions for responsible, innovation-friendly AI.

Importantly, the AI Act's transparency obligations could exert market pressure on developers to negotiate licenses rather than rely on the TDM exception as a legal loophole. Public disclosure of training sources may expose reputational risks for platforms that fail to engage with creator communities, incentivizing voluntary agreements.[262] Yet transparency alone cannot substitute for a calibrated remuneration framework that ensures a fair share of value flows back to the creative sector.

While mandatory, fee-free TDM has undoubtedly accelerated data-driven research and innovation, it has also exposed a blind spot in the EU's copyright architecture: authors currently enjoy neither a right to negotiate nor a right to payment when their works serve as the building blocks of generative AI. The policy options outlined above—and further elaborated in Chapter 4—seek to bridge this gap by combining industry-led licensing initiatives, collective bargaining mechanisms, and, if necessary, new statutory entitlements. The ultimate goal is to preserve the dynamism of AI development while safeguarding the economic and moral interests of the creators whose ingenuity underlies Europe's rich cultural heritage.

Beyond questions of legal design and economic efficiency, this debate ultimately touches upon the foundational values of the copyright system. The large-scale, uncompensated use of human literary and artistic works in AI training risks eroding the right to fair remuneration—an essential mechanism for sustaining creative labour in the digital era.[263] Fair compensation is not only a matter of distributive justice, but also of safeguarding the long-term vitality of human expression, including the forms of creativity that may eventually be enhanced through AI-assisted tools.[264] Unlike automated outputs generated by machine learning models, human literary and artistic works perform a unique cultural

[262] See OECD, Intellectual Property Issues in Artificial Intelligence Trained on Scraped Data, cit., at 18 (noting that a lack of transparency about dataset provenance hampers rights holders' ability to verify use and enforce their right).

[263] See e.g. Martin Senftleben, Generative AI and Author Remuneration. 54 IIC – International Review of Intellectual Property and Competition Law, 1535 (2023); Giancarlo Frosio, Should We Ban Generative AI, Incentivise It or Make It a Medium for Inclusive Creativity?" in E Bonadio and C Sganga (eds), *A Research Agenda for EU Copyright Law* 61 (Cheltenham, Edward Elgar, 2025) (arguing that generative AI risks parasitically exploiting human creativity and undermining the distinct social and cultural value of human authorship).

[264] See Authors', Performers' and Other Creative Workers' Organisations Joint Statement on Artificial Intelligence and the Draft AI Act (2023), Available at https://europeanwriterscouncil.eu/wp-content/uploads/2023/09/1414-authors-performers-and-other-creative-workers-organisations-joint.pdf

function.[265] They serve as a reflection of individual and collective identities, contributing to democratic dialogue and societal cohesion in ways that generative systems cannot replicate.

2.5. The AI Act and transparency obligations

The EU's AI Act introduces, for the first time, a requirement for providers of general-purpose AI models to "draw up and make publicly available a sufficiently detailed summary about the content used for training of the general-purpose AI model, according to a template provided by the AI Office".[266] This provision, in principle, aims to empower rightsholders by enabling them to verify whether their works have been used without authorization during the training of generative AI systems. However, a closer analysis reveals that this approach is structurally inadequate and fails to meaningfully address the real obstacles faced by individual creators.

The core weakness stems from the fact that training data transparency requirements are being layered on top of a fundamentally flawed legal foundation, namely the Article 4(3) opt-out mechanism of the CDSM Directive.[267] As highlighted in the legislative history, the opt-out was already affected by profound logistical challenges: no standardized machine-readable opt-out exists, no central registry of opted-out works is available, and the burden remains entirely on individual authors to monitor and enforce their rights.[268] Far from solving these issues, the AI Act merely assumes that a summary of training data will enable rightsholders to vindicate their rights—an assumption that collapses under practical scrutiny.

Firstly, the AI Act requires only a "sufficiently detailed summary" of training data—not the disclosure of the data itself. As discussed, given the immense scale and heterogeneity of modern AI training datasets, such summaries are almost certain to be incomplete, vague, and effectively useless for identifying specific unauthorized uses. The emphasis placed in Recital 107 on protecting trade secrets

---

[265] See e.g. Christophe Geiger, Building an Ethical Framework for Intellectual Property in the EU: Time to Revise the Charter of Fundamental Rights, in G. Ghidini and V. Falce (eds), Reforming Intellectual Property 77 (Edward Elgar, 2022) (contrasting human creativity with mere reproduction or economic exploitation, and arguing that protection should reflect the "moral and cultural values" underpinning society).

[266] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139, and (EU) 2019/2144, and Directives 2014/90/EU, (EU) 2016/797, and (EU) 2020/1828 (Artificial Intelligence Act), OJ L 1689, 12.7.2024, p. 1–144 (hereinafter: EU AI ACT).

[267] See Martin Kretschmer, et al., Copyright Law and the Lifecycle of Machine Learning Models, 55 IIC - International Review of Intellectual Property and Competition Law 110 (2024) (pointing out that requiring disclosure of training data operationalizes the opt-out mechanism but does not create new exceptions or rights, only enforces compliance)

[268] See Martin Senftleben, The TDM Opt-Out in the EU – Five Problems, One Solution, Kluwer Copyright Blog (April 22, 2025). Available at https://copyrightblog.kluweriplaw.com/2025/04/22/the-tdm-opt-out-in-the-eu-five-problems-one-solution/ (observing that Article 4, which has become central to the regulation of commercial AI training activities in the EU, was added only at the final stages of the legislative process, without a comprehensive impact assessment of its implications for the development of generative AI systems (GenAI) and the protection of authors' and rightsholders' interests)See also Thomas Margoni and Martin Kretschmer, A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology, 71 *GRUR International*, 685-701 at 688-90 (2022), (noting that the requirement of lawful access is difficult to operationalize... leading to practical obstacles for those wishing to rely on the exception).

and confidential information further narrows the scope of disclosure, ensuring that rightsholders will not receive the granularity of information needed to verify whether their works have been included.[269] This limitation reflects a broader regulatory tension: while the AI Act promotes transparency through disclosure obligations, it must also respect the protections granted to confidential business information under the Trade Secrets Directive (Directive (EU) 2016/943). As a result, GPAI providers may lawfully withhold information that could expose proprietary training datasets—especially where such datasets derive their value from exclusivity and have been subject to reasonable confidentiality measures.

Second, while Article 53(1)(c) of the AI Act[270] compels AI providers to **implement a policy to comply with Union copyright law**, including the **requirement to respect opt-outs expressed under Article 4(3) of the CDSM Directive**, the **policy-based approach** has severe limitations.[271] This policy obligation merely formalizes what was already a theoretical legal duty. It does not create enforceable obligations capable of ensuring meaningful compliance, especially against **providers located outside the EU**, who dominate the development of general-purpose AI models.[272] Although the Act tries to address this by extending obligations to **any provider placing a GPAI model on the EU market**, the extraterritorial enforcement of these obligations remains highly uncertain due to the territorial nature of copyright law and the practical difficulty of pursuing infringement claims when training occurs under foreign legal standards, such as U.S. fair use.[273] Without robust audit rights, penalties, or automatic enforcement mechanisms, the AI Act relies entirely on providers' goodwill and voluntary compliance—an approach that, as past behavior by major AI developers shows, is naïve at best.[274] In addition, the exact meaning of the references to "Union law on copyright and related rights" and "Union copyright law" remains uncertain, because there is no copyright equivalent of the unitary EU trade mark or design: the exclusive rights granted under copyright law continue to be national in scope and apply only within the territory of each Member State.[275]

---

[269] See Alexander Peukert, Copyright in the Artificial Intelligence Act – A Primer, 73 GRUR International 497, 502 (2024); Adam, Buick, Copyright and AI training data—transparency to the rescue? 20 Journal of Intellectual Property Law & Practice, 182, 190 (2025).

[270] See Art. 53(1)(d) of the EU AI Act.

[271] On whether Article 53(1)(c) AI Act implies that the training of generative AI models is covered by the TDM exception in Article 4 CDSM, see Section 2.1.3. As discussed there, the AI Act does not expand or clarify the substantive scope of EU copyright exceptions. Rather, it presupposes compliance where the TDM exception is validly applicable, without adjudicating the legality of AI training itself under Article 4 CDSM.

[272] See Adam, Buick, Copyright and AI training data—transparency to the rescue?, cit. at 190-191.

[273] See also Directive 2004/48/EC ("Enforcement Directive"), which provides the general legal framework for enforcing intellectual property rights in the EU. However, its practical applicability to non-EU GPAI providers remains uncertain.

[274] See, e.g., recent agreements signed by OpenAI with major media outlets such as Axel Springer, The Financial Times, Le Monde, and Associated Press, allowing the company to access and license their copyrighted content for training purposes (supra note 137). While these agreements signal a shift toward negotiated use, they also implicitly acknowledge that past ingestion practices likely lacked adequate authorisation. These developments underscore the inadequacy of relying on voluntary compliance in the absence of enforceable legal obligations.

[275] See Alexander Peukert, Copyright in the Artificial Intelligence Act – A Primer, cit. at 504 ((noting that, despite significant harmonization, copyright remains nationally based within the EU, and that it is unclear whether the obligation to comply with "Union copyright law" under Article 53(1)(c) AI Act refers to national laws as harmonized collectively or only to directly harmonized elements).

Thirdly, the AI Act grossly underestimates the problems associated with extraterritorial application. While it formally imposes obligations on any GPAI model "placed on the market" in the EU, this provision may be insufficient to deter companies from relocating training activities abroad, further weakening Europe's strategic position in the global AI race.[276] As commentators have warned, this dynamic is likely to encourage relocation of training pipelines outside EU borders, further exacerbating Europe's already precarious position in the global AI race.[277] The idea that a mere summary of training data can bridge the enormous gap between different copyright regimes is, frankly, untenable.

Moreover, the recently drafted General-Purpose AI Code of Practice only amplifies these doubts.[278] While it outlines machine-readable opt-outs, copyright policies, and complaint mechanisms, these are built on vague standards, voluntary participation, and no enforcement mechanisms. Without legal consequences for non-compliance, the incentives for providers to meaningfully engage with these commitments—especially when compliance may increase their exposure to litigation—are almost non-existent.

The AI Act does not address the underlying structural challenge: individual clearance at scale is not realistically feasible. Relying on transparency alone to facilitate a functioning rights market is, at best, an overly optimistic assumption. Transaction costs would likely overwhelm any such system, and there is currently no viable pathway for developing collective rights management structures or automated licensing mechanisms capable of operationalising rightsholder entitlements at scale. Even proposals for automated licensing solutions—such as using bots to detect machine-readable opt-outs—encounter serious limitations. In particular, the challenge of reliably verifying the identity of rightsholders in a decentralised, global information environment remains unresolved.[279]

Thus, it is already evident that reliance on transparency requirements—supplemented by a general obligation to implement a policy respecting "Union copyright law"—is insufficient to achieve the presumed objective of ensuring that individual authors are fairly compensated for the use of their works in AI training data. A more credible strategy would have required rethinking the underlying copyright infrastructure—through the introduction of statutory collective licensing models tailored to AI training or the development of EU-level centralized rights management platforms capable of handling opt-outs and licensing requests at scale. While complex, such reforms would be necessary to materially improve the position of authors, rather than offering only symbolic recognition of their concerns. A forward-looking framework could combine elements of mandatory collective rights management, robust public oversight, and genuine opt-in mechanisms. It would shift the burden away from individual authors and

[276] See Adam, Buick, Copyright and AI training data—transparency to the rescue?, cit. at 191.

[277] See Thomas Margoni & Martin Kretschmer, A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology, 71 GRUR Int'l 685 (2022).

[278] See European Commission, Working Groups of the First General-Purpose AI Code of Practice, Third Draft of the General-Purpose AI Code of Practice – Copyright Section, April 2024, available at: https://ec.europa.eu/newsroom/dae/redirection/document/113606

[279] See e.g. Paul Keller, Considerations for Opt-Out Compliance Policies by AI Model Developers, Open Future, May 16, 2024, https://openfuture.eu/wp-content/uploads/2024/05/240516considerations_of_opt-out_compliance_policies.pdf; Martin, Senftleben, AI Act and Author Remuneration - A Model for Other Regions? (February 24, 2024). Available at SSRN: https://ssrn.com/abstract=4740268

toward a system designed to proactively safeguard their rights, while enabling responsible AI development under clear, predictable, and equitable rules.

While the preceding discussion highlights structural limitations in the EU's current copyright framework, it is also important to assess the practical impact of the AI Act's existing transparency provisions. Article 53(1)(c)-(d) establishes binding legal duties: GPAI providers must develop a copyright compliance policy and publish detailed summaries of the datasets used to train their models, following a Commission-defined template. These obligations mark an important step forward compared to previous voluntary frameworks. However, their effectiveness will depend heavily on the quality of enforcement and the specificity of forthcoming implementing acts. Without clear guidance on what constitutes an adequate summary or policy, and in the absence of robust audit mechanisms, there is a risk that compliance may become formalistic or superficial. While these measures are not merely symbolic in legal terms, their ability to ensure meaningful accountability remains to be tested in practice. It is important to stress, however, that the transparency obligations stipulated in Article 53—requiring copyright compliance policies, respect for opt-outs under the DSM Directive, and publication of training data summaries—are procedural in nature. While they aim to enhance transparency and facilitate enforcement, they do not modify the substantive scope of existing copyright exceptions or introduce new rights clearance mechanisms.[280] This limitation underscores the need for complementary reforms that address the underlying legal and economic asymmetries in AI training practices. These concerns are particularly problematic for open-source GPAI models, which often lack a centralised organisational structure and are developed collaboratively by research groups or community-led projects. While Article 53 formally applies to all GPAI providers, including those releasing models under free or open-source licences, the decentralized nature of these initiatives raises serious questions about enforcement and compliance. In many cases, it is unclear who the 'provider' is for the purposes of Article 53 obligations—especially when models are forked or maintained by informal collectives. Moreover, the documentation of training data in open-source projects is frequently incomplete or inconsistent, complicating efforts to produce the "sufficiently detailed summary" required by the AI Act. These structural challenges do not negate the importance of transparency, but they do suggest that a more proportionate and risk-sensitive implementation is necessary. Section 4.3(k) of this study offers specific policy recommendations to address this issue, including tailored compliance thresholds and modular disclosure templates.

The AI Act then presupposes that developers operate within the existing legal framework, without itself creating new exceptions or authorizing activities beyond what the DSM Directive permits. Therefore, if the training of generative AI models is found to fall outside the TDM exceptions, the AI Act does not retroactively legalize such activities. It merely regulates compliance and disclosure. **In this context, it is also worth recalling that existing copyright safeguards—particularly Article 19 of the CDSM Directive—could play a complementary role in reinforcing transparency obligations under the AI Act.** Article 19 entitles authors and performers to receive regular, comprehensive information about the exploitation of their works, including all revenues generated and remuneration due. This obligation

[280] On this, see See Martin Kretschmer, et al., Copyright Law and the Lifecycle of Machine Learning Models, cit.

extends beyond direct contractual partners to sub-licensees, where necessary, and is explicitly protected from contractual override under Article 23. By contrast to the AI Act's vague and procedural disclosure model, Article 19 offers a substantive, enforceable right to information that could be analogically extended to the use of protected works in AI training contexts. If interpreted coherently, this existing EU copyright mechanism could support a more robust enforcement of training data transparency—ensuring that creators are not left to rely solely on vague summaries or provider goodwill. In this light, future implementation of the AI Act should take into account the normative weight and enforceability of these copyright-specific transparency obligations.

It is therefore clear that neither the AI Act nor the existing copyright framework—despite offering important procedural safeguards—resolves the core legal uncertainty surrounding generative AI training. Article 53 of the AI Act mandates disclosure and compliance policies, and Articles 18 and 19 of the CDSM Directive reinforce transparency and fair remuneration. Yet these provisions operate within a legal system that has not yet determined whether the large-scale ingestion of protected works by AI systems is lawful in the first place. Until this substantive question is clarified—either through CJEU interpretation or legislative reform—transparency alone cannot compensate for structural asymmetries in bargaining power, enforcement, or access to redress. Legal coherence, not procedural layering, remains the central missing piece.

Table 4: What is the AI Act doing?

ASPECT	WHAT THE AI ACT DOES	WHAT IT DOES NOT DO
Regulation of behavior	It regulates how GPAI providers must behave: adopt copyright policies, respect opt-outs, and publish summaries.	It does not create new copyright exceptions or legalise acts that would otherwise infringe under EU law.
Acknowledgment of practice	It acknowledges that TDM techniques are used in AI training and introduces procedural safeguards.	It does not determine whether such use falls within the scope of Articles 3 or 4 of the CDSM Directive.
Compliance framework	It introduces a procedural framework (e.g. dataset summaries, Codes of Practice) for transparency and oversight.	It does not resolve the legal uncertainty about whether generative AI training is lawful under current EU rules.

3. LEGAL STATUS OF AI-GENERATED OUTPUTS (OUTPUT SIDE)

KEY FINDINGS:

Human Authorship is Central: EU copyright law only protects works that are the result of a human's intellectual creation. Fully autonomous AI-generated outputs—without meaningful human input—are excluded from protection.

AI-Assisted vs. AI-Generated: Legal eligibility depends on the degree of human involvement. AI tools used under human creative control may lead to protectable works; outputs created by AI alone do not.

No Copyright for Prompts Alone: Merely providing a prompt to an AI model does not amount to authorship. Human contributions must shape the expressive aspects of the output.

Public Domain by Default: Outputs with no human authorship fall into the public domain. This promotes openness but may undermine investment and raise competition issues.

No Legal Recognition of AI as Author: Unlike UK or business-oriented approaches, EU law rejects the concept of non-human or legal person authorship.

Risk of Infringement Still Applies: Even if AI-generated content isn't protectable, it may still infringe existing copyrights—especially when outputs reproduce or resemble training data.

Emerging Policy Options: Scholars propose tiered authorship models or sui generis rights to address grey areas, but legislative reform remains politically and conceptually difficult.

Liability can be attributed to individuals or legal entities: Users or providers may face liability if AI outputs unlawfully reuse protected content. Compliance with the AI Act and copyright laws is essential.

Creative Control Is the Threshold: Copyright protection hinges not on the use of AI, but on whether the human made free and creative choices that shaped the final output. Courts must assess the depth and impact of human involvement.

Style Is Not Protected, But Risks Remain: Imitating an artist's style (e.g., "in the style of Van Gogh") is not copyright infringement, but raises fairness and reputational concerns. These may fall outside copyright law but could implicate unfair competition.

No General Exception for AI Outputs: EU law does not provide any general exception for AI-generated outputs that infringe third-party rights. Outputs that reproduce protected material remain unlawful without a valid exception.

Generative AI systems produce new content by learning patterns from large datasets. These systems rely on techniques such as deep learning and neural networks to synthesise original-seeming material that is often indistinguishable from human-created works. Unlike traditional software tools, which follow rule-based instructions, GenAI models operate through probabilistic reasoning and data-driven generalisations. As such, they do not merely retrieve or remix existing content, but produce statistically derived outputs that resemble new forms of expression, modelled on prior data exposure. This fundamental shift in machine capability—from automation to generation—raises other questions for intellectual property frameworks, particularly in relation to authorship, ownership, and originality.

Despite the rapid evolution of generative technologies, the prevailing consensus in both European and international legal systems is that copyright protection remains fundamentally tied to human authorship.[281] Jurisdictions such as the European Union (EU), United States, and China[282] currently exclude fully AI-processed outputs from copyright protection when no meaningful human input is identifiable. This reinforces the anthropocentric structure of IP systems, which are built on the premise that creative expression is a uniquely human attribute and that legal authorship must be traceable to a natural person.

The resulting legal uncertainty has triggered a growing debate about whether current legal tools are fit for purpose. Some experts and stakeholders advocate for the development of sui generis rights[283] or other alternative mechanisms to fill the perceived gap in protection for AI-generated content.[284] Others emphasise the importance of preserving the public domain, warning that expanding IP rights to non-human outputs could distort incentive structures, exacerbate market concentration, and reduce access to cultural and creative resources.

This section examines the legal issues raised by AI-generated content under EU law, focusing on originality, authorship, and the distinction between AI-assisted versus fully AI-generated outputs.

3.1. Originality and authorship under EU law

If a poem or painting is synthetically produced by an AI system through automated processing, can any person lawfully claim it as their intellectual property? This seemingly simple question lies at the center of a deep legal and philosophical debate that has intensified in recent years, as the outputs of generative artificial intelligence (AI) systems begin to resemble the creative works traditionally protected by copyright law. Imagine a museum exhibition showcasing images generated entirely by an AI using prompts like "a moonlit forest in the style of Van Gogh." The human curator may have typed the prompt, but the intricate brushstrokes, composition, and texture were the work of an algorithm. Who, if anyone, owns this creation? Under current 'EU copyright law', the answer is unequivocal: no one. The EU's legal framework for copyright does not recognize non-human entities as authors, and

[281] See e.g. Jane Ginsburg and Luke Budiardjo, Authors and Machines, 34 Berkeley Technology Law Journal 343, 346 (2019) (arguing that both conception and execution are required elements of authorship, and that machine-generated outputs lacking human involvement in these stages fall outside copyright protection).

[282] China's Copyright Law recognises only natural persons and legal entities as authors; however, some Chinese courts have recognised the copyrightability of AI-generated works when they involve human intellectual activities and have considered the user of the AI software as the copyright owner. See Copyright Law of the People's Republic of China of Feb. 26, 2010, art 12; Yong Wan and Hongxuyang Lu,Copyright protection for AI-generated outputs: The experience from China, 42 Computer Law & Security Review (2021).

[283] See e.g., Enrico Bonadio & Luke McDonagh, Artificial Intelligence as Producer and Consumer of Copyright Works: Evaluating the Consequences of Algorithmic Creativity' Intellectual Property Quarterly 112-137 (2020) (proposing a thin sui generis right); Benjamin Hardman and James Housel, A Sui Generis Approach to the Protection of AI-Generated Works: Balancing Innovation and Authorship (August 30, 2023). Available at SSRN: https://ssrn.com/abstract=4557004; Ana Ramalho, Will Robots Rule the (Artistic) World? A Proposed Model for the Legal Status of Creations by Artificial Intelligence Systems", 21 Journal of Internet Law 1 (2017).

[284] See e.g., Jane Ginsburg and Luke Budiardjo, Authors and Machines, cit., at 445 (arguing that authorless outputs should not receive copyright protection and that expanding IP to cover such works could undermine existing frameworks unless clearly justified).

without human intellectual contribution, such outputs are not eligible for copyright protection. This section explores the legal foundations of this principle and examines its consequences for the evolving landscape of AI-driven creativity.

Human-Centric Copyright: The Author's Own Intellectual Creation

Under EU copyright doctrine, the concept of authorship is firmly rooted in the notion that authorship requires intentional, human-originated expression. The foundational requirement for copyright protection is that a work must be the "author's own intellectual creation"—a standard that has been shaped by a consistent line of jurisprudence from the Court of Justice of the European Union (CJEU). The seminal case Infopaq International A/S v. Danske Dagblades Forening (C-5/08)[285] laid down the principle that copyright subsists only in subject matter that reflects the "author's own intellectual creation," defined as the expression of the author's free and creative choices. This human-centric threshold has since been reaffirmed in several key decisions, including Painer (C-145/10),[286] Football Dataco (C-604/10),[287] and Levola Hengelo (C-310/17),[288] all of which stress the necessity of a human making creative decisions that stamp the work with their personal imprint. In particular, in the Infopaq decision the Court noted that "only through the choice, sequence and combination of those words" can an author express creativity in a manner that results in a protectable work. The idea–expression dichotomy, long central also to European copyright law, thus presupposes the presence of a human subject capable of making autonomous creative decisions. The principle was reaffirmed in Painer, where the CJEU held that even in media with limited expressive range—such as photography—copyright subsists if the author is able to imprint the work with a "personal touch" through choices like framing, lighting, or timing. The EU thus excludes machine-generated outputs irrespective of their apparent originality or aesthetic value but because they lack the personal imprint of a human author.[289] By contrast, where outputs are generated automatically by AI systems without such human intervention, there is no room for original expression. The result is that such outputs, however novel or convincing they may appear, fall outside the scope of protection under EU copyright law. As Advocate General Trstenjak underscored in Football Association Premier League, "only human creations are protected."[290]

Unlike jurisdictions such as the United Kingdom—which, in Section 9(3) of the Copyright, Designs and Patents Act 1988, allows for the attribution of authorship in computer-generated works to the person who made the arrangements necessary for their creation[291] —EU copyright law follows a different approach. It does not recognise authorship in the absence of human creativity, nor does it permit default attribution to non-human or legal persons. This difference reflects a deeper philosophical

[285] C-05/08, Infopaq International v. Danske Dagblades Forening (2009) ECLI:EU:C:2009:465 (Infopaq)

[286] C-145/10, Eva-Maria Painer v Standard VerlagsGmbH and Others, ECLI:EU:C:2011:798.

[287] C-604/10, Football Dataco Ltd and Others, ECLI:EU:C:2012:115.

[288] Case C-310/17, Levola Hengelo BV v. Smile Foods BV, ECLI:EU:C:2018:899.

[289] See Enrico Bonadio et. al., Will Technology-Aided Creativity Force Us to Rethink Copyright's Fundamentals? Highlights from the Platform Economy and Artificial Intelligence, cit., at 1188.

[290] See Opinion of Advocate General Trstenjak, Case C-145/10, Painer v. Standard VerlagsGmbH, 2011 E.C.R. I-12533, § 121 (Apr. 12, 2011).

[291] See § 9(3) of the UK Copyright, Designs and Patents Act 1988.

divergence: while some legal systems are willing to stretch the concept of authorship to cover the realities of machine-made creativity, the EU remains doctrinally committed to the idea that copyright is essentially anthropocentric.[292]

A similar contrast emerges in the United States. Although U.S. copyright law requires only a "modicum of creativity" for protection—following the Supreme Court's ruling in Feist Publications v. Rural Telephone Service (1991)[293]—it too maintains the requirement of human authorship. The U.S. Copyright Office has repeatedly clarified that works generated by AI systems, without direct human involvement, do not qualify for protection under existing copyright law.[294] Thus, despite different thresholds for originality, both EU and U.S. law converge in excluding non-human creations from the scope of copyright.

This convergence reflects deeper structural and conceptual similarities between the two systems. In the U.S., as in the EU, the very notion of copyrightable subject matter rests on the assumption that authorship is inherently human. Core concepts such as authorship, originality, and the expression of ideas all presuppose a human agent making creative choices. This means that without a human author, there is no "expression" in the legal sense—no transformation of ideas into protectable form. Rather, the product of an autonomous system remains outside the legal definition of a "work." U.S. case law, including the foundational Burrow-Giles Lithographic Co. v. Sarony,[295] reinforces this view by tying protection to the author's intellectual conception and execution of the work. Consequently, proposals to grant copyright to AI-generated outputs would require not just legislative adjustment, but a fundamental rethinking of the system's normative underpinnings.

Implications for AI-Generated Outputs

Given this framework, it follows that purely AI-generated outputs—those created automatically by an AI system without substantial human intervention—are not eligible for copyright protection in the EU. Such outputs are considered to fall into the public domain, making them freely available for anyone to use, reproduce, or adapt without seeking permission or providing attribution. The legal and commercial implications of this are significant. For creators and companies investing in AI systems that generate music, art, or text, there is no proprietary right over the final output unless a human has contributed in a way that meets the "intellectual creation" standard.

---

[292] See Enrico Bonadio et. al., Will Technology-Aided Creativity Force Us to Rethink Copyright's Fundamentals? Highlights from the Platform Economy and Artificial Intelligence, cit., at 1188.

[293] See Feist Publications, Inc., v. Rural Telephone Service Co., 499 U.S. 340 (1991).

[294] See United States Copyright Office, Compendium of U.S. Copyright Office Practices (3d ed. 2021) §313.2; Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence, 88 Fed. Reg. 16190, 16192 (Mar. 16, 2023); U.S. Copyright Office, Copyright and Artificial Intelligence: Part 2: Copyrightability (2025) available at https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-2-Copyrightability-Report.pdf

[295] See Burrow-Giles Lithographic Co. v. Sarony, 111 U.S. 53 (1884).

The practical application of this principle was made explicit in a recent Czech court ruling from 2023, which has since become a reference point in European debates around AI authorship.[296] In this case, the court addressed whether an image generated by an AI platform—prompted by a user who entered a detailed textual description—could be protected by copyright. The court concluded that the human's contribution in writing the prompt did not amount to authorship under copyright law. Since the human operator had not made any creative choices in the expressive form of the image (e.g., composition, colour, shading), and the AI system had assembled the output based on its training data and internal rules, the work was not considered eligible for protection. Therefore, prompting can be seen as more akin to generating ideas than expressions. This judgment affirms the EU position that simply operating an AI tool, or providing an idea or input, does not suffice to establish authorship if the creative expression is determined by the system itself.[297]

This position has been reinforced by the European Commission's 2020 report, "Trends and Developments in Artificial Intelligence – Challenges to the Intellectual Property Rights Framework", which emphasizes that existing EU copyright law "requires a human author" and that "fully autonomous AI-processed outputs currently fall outside the scope of copyright protection".[298] The report further notes that related rights or sui generis mechanisms might eventually be explored for such works, but for now, the legal framework remains centered on the human creator.

This legal model also highlights the growing difficulty in assessing the threshold of human involvement in AI-assisted creativity. If a user simply prompts a system with a general instruction (e.g. "Compose a poem of spiritual and allegorical nature in the style of Dante") and accepts the first output without further modification, their contribution is unlikely to satisfy the standard of originality. By contrast, if the user iteratively refines the result, edits the output, or integrates it into a broader creative work, their role may be deemed sufficiently creative to justify authorship. The challenge for policymakers and courts is to develop clear criteria to distinguish between mere use of a tool and substantive human authorship. As mentioned, proposals such as tiered authorship models or sui generis rights have emerged to address these grey areas.[299] However, they require complex legal reform and risk upsetting the balance between IP protection and the public domain. In addition, extending rights to machine-generated content may concentrate power in the hands of platform owners and dilute the concept of

[296] Rozsudek Městského soudu v Praze z 11.října sp. zn. 10 C 13/2023. Available at https://justice.cz/documents/14569/1865919/10C_13_2023_10/108cad3e-d9e8-454f-bfac-d58e1253c83a (Decision of the Municipal Court Prague from 11. October, no 10 C 13/2023).

[297] See European Commission: Directorate-General for Communications Networks, Content and Technology, Hartmann, C. et al., Trends and developments in artificial intelligence – Challenges to the intellectual property rights framework – Final report, Publications Office of the European Union, 2020, at 116. Available at https://data.europa.eu/doi/10.2759/683128

[298] Ibidem.

[299] See e.g. Enrico Bonadio & Luke McDonagh, 'Artificial Intelligence as Producer and Consumer of Copyright Works: Evaluating the Consequences of Algorithmic Creativity' Intellectual Property Quarterly 112-137 (2020) (proposing that works generated by AI should not receive full copyright protection but could instead be covered by a thin sui generis right); Benjamin Hardman and James Housel, A Sui Generis Approach to the Protection of AI-Generated Works: Balancing Innovation and Authorship (August 30, 2023). Available at SSRN: https://ssrn.com/abstract=4557004; Ana Ramalho, Will Robots Rule the (Artistic) World? A Proposed Model for the Legal Status of Creations by Artificial Intelligence Systems", 21 Journal of Internet Law 1 (2017).

human authorship. This raises a deeper conceptual and legal question: how should law distinguish between meaningful creative intervention and mere tool usage in the context of AI?

Beyond the Prompt: Where Does Human Creativity End?

The central issue, therefore, becomes where to draw the line between AI-assisted and AI-processed outputs. If a human uses AI as a tool—much like a brush or a camera—while making substantive creative decisions, the resulting output may still qualify for copyright protection. This aligns with the logic of the CJEU's reasoning: it is not the use of technology per se that disqualifies a work, but the absence of identifiable human intellectual input.

For instance, if a graphic designer uses AI to generate background patterns and then integrates, edits, and transforms these elements into a larger composition with their own creative decisions, the resulting work may reflect sufficient human authorship. But if the designer merely inputs a textual prompt into a generative model and accepts the first image output without further intervention or modification, the situation becomes more legally uncertain. Courts and policymakers must therefore grapple with questions of degree, threshold, and intent. Key factors in this assessment may include the degree of creative control exercised by the human, the extent to which the output reflects identifiable personal choices, and whether the human contribution involves the selection, arrangement, or meaningful transformation of AI-generated material. In light of the difficulty in drawing a clear boundary between human and machine creativity, recent academic literature has proposed different models to operationalize this distinction. Some authors suggest the adoption of a tiered framework distinguishing between AI-assisted and AI-processed outputs, where the former may still benefit from copyright protection if human creative input can be clearly identified and documented.[300] Others – as already seen - advocate for *a sui generis or thin* right tailored to protect investments in AI-generated content without invoking traditional authorship criteria.[301] However, such reform would not only require complex legislative action at the EU level and consensus among Member States—which remains unlikely in the near term—but would also presuppose the need to completely modify the anthropocentric foundation of copyright law. Given the enduring normative and cultural significance of human authorship in European legal tradition, it is far from clear that this shift is either desirable or necessary.

The Human Element as the Legal Bedrock

[300] See e.g. Vincenzo Iaia, To Be, or Not to Be ... Original Under Copyright Law, That Is (One of) the Main Questions Concerning AI-Produced Works, 71 GRUR International, 793–812 (2022); Peter Mezei, "You Ain't Seen Nothing Yet" - Arguments against the Protectability of AI-generated Outputs by Copyright Law. In: Maurizio Borghi - Roger Brownsword (eds.): Informational Rights and Informational Wrongs: A Tapestry for Our Times, 126-143 (Routledge 2023); Benjamin Hardman and James Housel, A Sui Generis Approach to the Protection of AI-Generated Works: Balancing Innovation and Authorship (August 30, 2023). Available at SSRN: https://ssrn.com/abstract=4557004.

[301] See e.g. Enrico Bonadio & Luke McDonagh, Artificial Intelligence as Producer and Consumer of Copyright Works: Evaluating the Consequences of Algorithmic Creativity, Intellectual Property Quarterly 2, 112-137 (2020); Ana Ramalho, Will Robots Rule the (Artistic) World? A Proposed Model for the Legal Status of Creations by Artificial Intelligence Systems", 21 Journal of Internet Law 1 (2017); Haochen Sun, Redesigning Copyright Protection in the Era of Artificial Intelligence, 107 Iowa L. Rev. 1213 (2022); Anne Lauber-Rönsberg and Sven Hetmank, The concept of authorship and inventorship under pressure: Does artificial intelligence shift paradigms? 14 Journal of Intellectual Property Law & Practice 570–579 (2019); Benjamin Hardman and James Housel, A Sui Generis Approach to the Protection of AI-Generated Works, cit.

The EU copyright framework currently offers no protection for works generated entirely by AI in the absence of meaningful human creative input. The threshold for authorship remains tied to the expression of free and creative choices by a human author. While this doctrinal clarity provides legal certainty, it also reveals an emerging misalignment with technological developments. As generative AI systems become more autonomous and capable, policymakers may need to assess whether the existing framework remains adequate—or whether complementary legal tools, such as registration-based rights, neighbouring rights, or other regimes, are required to capture the economic value of non-human outputs without eroding the philosophical core of copyright law.

The unresolved tension between human authorship and machine-generated creativity is likely to intensify, particularly as large language models and generative image systems evolve. For now, however, the EU maintains a human-centered conception of intellectual property—a stance that reflects a continued belief in the unique value of human imagination, discernment, and responsibility in the creative process.

3.1.1. Does the Human-Centric Approach Still Make Sense in the Era of Advanced Generative AI?

The EU's insistence on human authorship as the cornerstone of copyright protection has so far provided a clear doctrinal anchor in a fast-changing technological environment. But as generative AI models evolve from narrow tools into increasingly autonomous systems—capable of producing complex creative outputs with minimal or no human intervention—the question becomes whether this anthropocentric legal model remains conceptually sound and practically viable. Put differently: should copyright protection remain exclusively tied to human intellectual input in a world where machines may soon exhibit behaviours that, to all appearances, mirror creativity? The European Parliament addressed this dilemma in its 2020 resolution, explicitly rejecting the idea of granting legal personality to AI systems and reaffirming that copyright protection should remain anchored in human intellectual creation.[302]

In line with this position, current EU law (as explained above) recognises copyright only in works reflecting an author's own intellectual creation. This entails free and creative human choices, not merely mechanical or algorithmic processes. As the CJEU has consistently reaffirmed, creativity must be linked to personal expression. This makes sense in a historical context where only humans could author works and where copyright was primarily justified by moral rights (dignity of the author) and utilitarian considerations (incentives for human innovation).

That said, the pace of technological change has increased substantially and shows no sign of slowing. Large language models, generative adversarial networks, and multi-modal AI agents are increasingly capable of producing novel, contextually rich, and stylistically coherent outputs that are often indistinguishable from human creations. AI-generated music, visual art, fiction, and even academic

[302] See European Parliament resolution of 20 October 2020 on intellectual property rights for the development of artificial intelligence technologies, 2020/2015(INI), 2021 O.J. (C 404) 129, at § 13.

writing are now mainstream phenomena. These systems can synthesize and recombine data in ways that exhibit what Boden calls "combinational" or even "transformational" creativity.[303]

The underlying code is not simply executing pre-written instructions but interacting probabilistically with vast corpora of data to generate seemingly "original" content. This trend complicates the binary distinction between human and machine agency.

From a legal standpoint, continuing to exclude all non-human outputs from copyright may create a growing mismatch between legal norms and social or economic practices. As noted in the European Commission's Trends and Developments in AI report, the use of AI in cultural production is becoming pervasive, especially in sectors like journalism, design, music, and pharmaceutical research.[304] In particular, the report affirms that AI-assisted works involving meaningful human input are still covered under EU law, but explicitly acknowledges that fully autonomous AI outputs fall outside existing copyright regimes.[305]

The current exclusion of non-human outputs from protection raises several emerging normative and policy concerns:

- **Market Distortion and Incentive Gaps**: As more valuable content is created without human authorship, the absence of IP protection could deter investment in AI creative systems, leading to under-incentivization. Without exclusive rights, companies might rely heavily on trade secrets or technological protection measures, which could limit access and openness.

- **Public Domain Saturation:** The uncontrolled proliferation of high-quality, unprotected content may erode the distinctiveness and economic value of traditionally authored works. If AI content floods the market, authors may find it harder to compete, both in visibility and in licensing value.

- **Authorship Attribution and Legal Ambiguity**: Even where humans are involved, the threshold for authorship becomes increasingly ambiguous. How many decisions must a human make to "own" an AI output? Prompt engineering, iterative curation, and fine-tuning of models may involve substantial expertise—should these acts be treated as acts of creation or as technical manipulation?

- **Ownership and Liability in Autonomous Systems**: As we move towards more autonomous AI agents—capable not only of creating but of initiating tasks, selecting data inputs, and refining outputs—the question of who should be accountable (and rewarded) for the work becomes more pressing. If the human role becomes so attenuated that it no longer meets the current threshold of "intellectual

[303] See Margaret A. Boden, *The Creative Mind: Myths and Mechanisms* (2nd ed., London: Routledge, 2004) (identifying three distinct types of creativity: combinational, exploratory, and transformational creativity. Combinational creativity involves the novel combination of familiar ideas. Exploratory creativity refers to the process of navigating a given conceptual space to generate new ideas within an established framework. Transformational creativity, the most radical form, entails modifying or fundamentally reshaping the conceptual space itself, thus enabling the emergence of previously inconceivable ideas). On this discussion, see also See Giorgio Franceschelli and Mirco Musolesi, On the creativity of large language models. AI & Soc 1, 3 (2024).

[304] See European Commission: Directorate-General for Communications Networks, Content and Technology, Hartmann, C. et al., Trends and developments in artificial intelligence – Challenges to the intellectual property rights framework – Final report, cit.

[305] Ibidem.

creation," yet the output is commercially or culturally valuable, a growing gap in the legal framework that may merit further evaluation as technologies evolve.

Despite these challenges, the current EU framework offers at least two important benefits that justify its continued relevance—at least in the medium term:

Normative Clarity: Anchoring copyright in human creativity aligns with the moral and philosophical foundations of European copyright doctrine, especially the emphasis on personality rights and the dignity of authorship. Recognizing machines as authors could dilute this framework and open the door to rights claims by corporations or platform owners without corresponding human expression.

Preservation of the Public Domain: By refusing to grant IP rights over purely machine-generated outputs, EU law avoids overreach. It ensures that the growing corpus of AI-generated content remains freely usable, promoting remix culture, innovation, and access. As noted by scholars like Mezei and Iaia, extending copyright to non-human agents risks "monopoly over abundance" and may undermine the balance between protection and the public domain.[306]

Still, this doctrinal integrity may not be enough in the long run. Several scholars and policymakers are now exploring intermediate or alternative approaches. For example, the already mentioned proposal for *a sui generis* right—distinct from copyright but offering limited-term protection for AI-generated outputs—is seen as an instrument to bridge the gap between incentive structures and doctrinal purity.[307] Such a right could be contingent on disclosure and registration, thereby enhancing transparency while avoiding the risk of blanket monopolies over machine-generated creativity.

Alternatively, a tiered authorship model, as discussed by Denicola and Frosio, could differentiate between outputs with high human involvement (AI-assisted) and those with minimal or no human input (fully AI-generated), allowing nuanced application of protection regimes.[308]

In such frameworks, rights might vest in the "creative director" or the entity that defined the parameters, not unlike the way film directors or software architects hold certain rights over collective works.

In conclusion, the EU's human-centric approach to copyright remains conceptually and normatively sound—particularly as a safeguard for the moral and cultural values deeply rooted in European legal traditions. While the future trajectory of generative AI and autonomous agents is uncertain, this unpredictability reinforces, rather than undermines, the value of a cautious and principled stance.

---

[306] See Vincenzo Iaia, To Be, or Not to Be ... Original Under Copyright Law, That Is (One of) the Main Questions Concerning AI-Produced Works, cit.; Peter Mezei, "You Ain't Seen Nothing Yet" - Arguments against the Protectability of AI-generated Outputs by Copyright Law, cit.

[307] See supra note 283.

[308] See Giancarlo Frosio, Four theories in search of an A(I)uthor, in Ryan Abbott (ed), Handbook of Artificial Intelligence and Intellectual Property 156 178 (Edward Elgar 2022) (arguing that a differentiated authorship framework recognizing varying levels of human input in AI outputs is necessary to align intellectual property protections with traditional copyright principles); Robert Denicola, *Ex Machina: Copyright Protection for Computer-Generated Works*, 69 Rutgers U. L. Rev. 251 (2016) (stressing the need to abandon rigid human authorship standards in favour of a spectrum-based approach that recognises varying degrees of human involvement in computer-generated outputs, thereby enabling a more nuanced application of copyright protection).

Maintaining the current framework as a normative anchor is not only defensible, but advisable. That said, to address the emerging grey zone between human-augmented and machine-driven creativity, the EU may need to refine how it operationalizes existing principles—developing nuanced legal tools that uphold the integrity of copyright without prematurely conceding to technological determinism. A legal framework fit for the AI age need not discard its human-centric foundation, but rather reaffirm it through careful, contextual adaptation.

Table 5: Copyright Eligibility of AI-Generated outputs under EU law

Type of Output	**Description of Human Involvement**	**Eligibility for Copyright Protection (EU)**	**Rationale / Legal Basis**
Human-Created Work	Entirely human-authored with no AI involvement	✅ Yes	Meets the originality requirement ('author's own intellectual creation')
AI-Assisted Work	Human uses AI as a tool; exercises creative control through prompting, editing, and integration into broader work	✅ Yes (case-by-case)	If human input reflects free and creative choices (Infopaq, Painer, C-310/17)
Prompt-Based Output with Minor Editing	Human enters detailed prompt and lightly edits AI-generated content	⚠ Uncertain	May fall short of 'personal imprint' threshold unless creative decisions are significant
Fully AI-processed outputs (Autonomous Output)	AI system generates content without meaningful human intervention or expressive contribution	❌ No	Fails originality threshold; lacks human authorship (EU law requires a natural person as author)
Corporate or Platform-Owned AI Output	AI is deployed by a company with no human creator identified	❌ No	No default attribution to legal entities under EU copyright (contrast with UK Section 9(3) CDPA 1988)

3.2. AI-assisted vs AI-generated: where to draw the line

The distinction between AI-assisted human works and fully AI-generated outputs is pivotal for maintaining coherence within the EU copyright framework. This conceptual division—acknowledged by

the European Parliament—reflects the differing regulatory challenges each category poses.[309] In simple terms, this distinction turns on the degree and nature of human creative involvement in the final expression. However, drawing a clear, operational line is increasingly challenging as AI tools become more sophisticated and integrated into creative processes.

At one end of the spectrum lie AI-assisted human works. Here, AI acts as a tool—often a highly sophisticated one—that supports but does not supplant human creativity. The human author exercises significant control over the expressive elements of the work, making creative choices that shape the final output. Examples include a photographer using AI-based enhancement software to adjust lighting conditions, or a writer employing a generative tool to produce a draft which they subsequently revise, rewrite, and refine extensively. In such cases, the human remains the principal creative agent: the AI merely facilitates or accelerates tasks that would otherwise be laborious. EU copyright law, grounded in the principle of originality as a manifestation of personal intellectual creation, is likely to recognise these outputs as human-authored works, provided that the human contribution is substantial and reflects the author's personal touch.

On the opposite end of the spectrum, we find outputs that are predominantly or entirely AI-generated. Here, human involvement is reduced to minimal, non-creative inputs—such as entering a simple prompt like "compose a poem about the rain" into a text generator and accepting the resulting output without meaningful modification. The creative expression itself—the choice and arrangement of words, the emotional tone, the stylistic nuances—is automatically generated by the AI system. Under prevailing legal doctrine in the EU, these outputs would not qualify for copyright protection, as they lack the requisite human creative input.

In practice, many creative workflows increasingly involve iterative human-AI collaboration, where human actors experiment with prompts, select from multiple outputs, provide feedback, and perform extensive post-processing.[310] Given this, for reasons both principled and pragmatic, the current approach—requiring significant human creative input for copyright protection—should be maintained and reinforced.

In order to provide greater clarity in practice, it may be helpful to differentiate between distinct categories of human interaction with generative systems. For example, a user may merely initiate the process by entering a prompt; in other cases, the user may iteratively refine outputs, select among variations, or make substantial post-editing contributions. There are also scenarios in which users blend AI-generated content with original material, creating hybrid works. The concept of "creative control" should therefore serve as a guiding interpretive tool. It requires courts and examiners to ask: Did the human author make free and creative choices that shaped the final expression in a meaningful way? If so, protection may attach—even if some elements were machine-generated.

[309] See European Parliament resolution of 20 October 2020 on intellectual property rights for the development of artificial intelligence technologies, 2020/2015(INI), 2021 O.J. (C 404) 129, at ∬ 14.

[310] See e.g Terrance Fong et al., A Survey of Socially Interactive Robots, 42 Robotics & Autonomous Systems 143,147/48 (2003) (observing that even minimal social cues in machines can lead users to perceive agency and intentional behaviour).

The threshold for protection in these cases should hinge on whether the final expression reflects the author's intellectual creation through genuine choices regarding structure, content, or style. Courts and policymakers may consider criteria such as the depth of intervention, the autonomy of the system, and the creative significance of the human role—not merely the fact of interaction—to assess authorship claims in AI-assisted works.

Recent academic and regulatory discussions have begun to converge around a functional distinction based on the degree of human input in AI-generated content. Notably, the U.S. Copyright Office's 2025 policy guidance reaffirms that copyright requires "sufficient human authorship," and excludes works generated without meaningful human creative control.[311] Building on this and similar analyses in legal scholarship,[312] a three-tiered model has emerged that distinguishes between: (i) outputs created with minimal human input (**generally unprotectable**), (ii) outputs shaped through meaningful human editing or curation (**potentially protectable**), and (iii) works in which AI is used purely as an auxiliary tool to support human authorship (**clearly protectable**). While this model is not codified, it reflects a growing consensus that could inform future administrative guidelines or soft-law instruments.

This taxonomy clarifies the conceptual boundaries between different types of AI involvement and provides a framework for assessing legal protection. It offers a lens through which to examine the normative and policy implications of extending—or withholding—copyright from non-human authored outputs. In what follows, three key arguments support maintaining the human-centric orientation of EU copyright law.

First, encouraging human creativity remains a fundamental normative goal. Granting exclusive rights to machine-generated outputs, absent meaningful human involvement, undermines the rationale for copyright: the promotion of human authorship and cultural enrichment. Non-human creativity, while impressive, does not align with the philosophical and constitutional justifications for IP protection within the European legal tradition. Copyright is not—and should not become—a system for rewarding machine activity.

Second, recognising copyright in AI-processed outputs would harm the public domain. Vast quantities of AI-generated material, lacking human authorship, could enter protected status—enclosing algorithmic recombination of existing cultural material. This could stifle innovation, restrict access to knowledge, and erode the commons. By contrast, treating non-human outputs as unprotected helps enrich the public domain. The public domain is a structurally necessary element of the EU copyright acquis, ensuring access to expired or unprotected content for reuse and development. Extending copyright to non-human outputs could undermine the proportionality and balance principles in Recital 3 of the InfoSoc Directive and CJEU jurisprudence. Thus, leaving AI-only outputs unprotected aligns with both normative and doctrinal principles.

[311] U.S. Copyright Office, Copyright and Artificial Intelligence: Part 2: Copyrightability (2025) available at https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-2-Copyrightability-Report.pdf

[312] See e.g. Johannes Fritz, Understanding authorship in Artificial Intelligence-assisted works, Journal of Intellectual Property Law & Practice, (2025).

Third, copyright protection for AI outputs would produce undesirable consequences. Ownership would likely vest in corporations that control AI systems, not individual creators. Companies could monopolise vast content without real human input, distorting the market and weakening copyright's credibility. It would also complicate enforcement: unclear authorship makes infringement, licensing, and moral rights difficult to assess.

For all these reasons, the distinction between AI-assisted and AI-generated outputs is foundational. Technological developments may blur operational boundaries, but the normative framework must insist on significant human input as a prerequisite for protection. The human must exert creative control over expressive choices, not merely initiate or supervise an automated process.

Accordingly, this study supports maintaining the human-centric approach. AI-assisted works should be assessed case by case, focusing on the extent and significance of human involvement. Courts and policymakers should resist blanket recognition of AI outputs. The analytical focus must remain on whether the final expression reflects personal intellectual creation—a standard embedded in EU copyright law.

Operationally, robust guidelines and best practices will be essential. This study proposes indicative (but non-exhaustive) criteria to inform assessment:

– the extent of human control over generation;

– the presence of creative choices in editing, structuring, or curation;

– the use of judgment in selecting or combining generated material;

– the degree of revision or refinement applied.

Prompting alone should not suffice. By contrast, when a user meaningfully shapes content through iterative engagement, aesthetic decisions, or integration with original elements, the output may meet the originality threshold. These criteria could inform future guidance from bodies like EUIPO or the AI Office. Where doubts persist, the presumption should favour the public domain, consistent with copyright's core principle of protecting human intellectual creation.

An additional concern arises when fully AI-generated content is presented as original despite lacking meaningful human input. Such outputs are ineligible for protection under EU law, but there is no systematic mechanism to verify or challenge false authorship claims. This could distort competition and mislead consumers. At scale, it could facilitate unfair practices, displacing authentic works. Transparency tools—like provenance tracing or metadata—may be needed, along with competition law scrutiny where market exclusion or dominance abuse is evident.

As AI becomes more pervasive in creative industries, maintaining the requirement of substantial human input remains essential to uphold copyright's goals. Safeguarding human authorship, preserving the public domain, and ensuring equitable participation in creative markets are not just technical matters—they are critical to a cultural ecosystem centred on human creativity.

That said, this does not preclude exploring complementary legal mechanisms tailored to AI-generated content. Such approaches must be carefully designed to avoid weakening originality standards or restricting access to the commons.

Beyond authorship, the designation of AI-generated outputs as public domain raises broader market concerns. While unprotected, these outputs can be monetised by dominant platforms with scale advantages, raising concerns over value distribution and sustainability for smaller actors. Creators lacking technical resources may be unable to monetise distinctive styles or maintain market presence. While style per se is not protected, its saturation by dominant players may trigger competition law scrutiny, especially if it leads to foreclosure or displacement of human creators. These concerns may merit attention from both copyright regulators and competition authorities.[313]

It is therefore clear that any reform should proceed cautiously, favouring solutions that enhance legal certainty and transparency without undermining the foundational values of EU copyright law. In this rapidly evolving landscape, it is these core principles—human creativity, proportionality, and openness—that must guide legal responses in the AI era.

3.3. Economic and Legal Challenges of AI-Generated Outputs: Disrupting Value Chains and Market Dynamics

While current policy discussions and legislative measures largely focused on the legal permissibility of using copyrighted content for training generative AI systems—particularly input-side exceptions such as Text and Data Mining (TDM),[314] a comprehensive regulatory framework must also confront the legal and economic implications of AI-generated outputs. This subsection briefly highlights the distinct challenges posed by such outputs, complementing the preceding analysis of authorship and originality. The EU copyright system was conceived at a time when technologies capable of algorithmically producing outputs that replicate or substitute human-created works did not yet exist.[315] Consequently, current rules are ill-equipped to respond to the economic and structural transformations triggered by generative AI. This technological leap disrupts traditional licensing models and challenges the foundational assumptions of the copyright system—namely, that creative outputs can be reliably attributed to identifiable human authors and managed through contractual or collective licensing schemes.

---

[313] See e.g. Giuseppe Colangelo, A Competition Policy Analysis of Copyright Protection in Gen AI, Singapore Journal of Legal Studies, forthcoming (2025). Available at https://ssrn.com/abstract=5201510 (arguing that interpreting copyright exceptions in light of market effects and competition policy could offer a more consistent and innovation-friendly framework for balancing access to data and incentives for human creativity in the GenAI context).

[314] See Articles 3 and 4 of the CDSM Directive.

[315] For detailed retrospective accounts of the legal and policy foundations of EU copyright and its evolution, see, e.g., P. Bernt Hugenholtz (ed.), The Future of Copyright in a Digital Environment (Kluwer 1996); Estelle Derclaye (ed.), Research Handbook on the Future of EU Copyright (Edward Elgar, 2009); Mireille van Eechoud et al., Harmonizing European Copyright Law: The Challenges of Better Lawmaking 57 (2009); Eleonora Rosati, Copyright and the Court of Justice of the European Union (Oxford University Press, 2023).

The proliferation of AI-generated content risks diluting the value of human authorship, complicating enforcement, and saturating markets with unattributed or unlicensed works. This dynamic weakens the competitive position of professional creators.[316] From a structural perspective, this may be understood as a market failure.[317] Automated content is produced faster, at lower cost, and at significantly higher volume—placing professional creators at a competitive disadvantage, not due to inferior quality but to structural conditions. Even robust licensing schemes may struggle to preserve the visibility and viability of human-created works within a digital ecosystem dominated by synthetic content. Stakeholders across the creative and cultural sectors have voiced growing concern that current rules fail to reflect the scope and consequences of this transformation. Collective management organisations, in particular, warn that AI-generated content may replicate and displace human-created works without adequate compensation mechanisms, further weakening the position of professional authors in already saturated distribution ecosystems.[318]

This challenge is increasingly recognised at the EU level. As seen, the AI Act introduces transparency obligations for general-purpose AI models, and Article 18 of the CDSM Directive affirms the principle of fair remuneration for authors. However, neither instrument directly addresses how the economic value derived from AI outputs should be distributed or regulated. Moreover, this regulatory blind spot is magnified by geopolitical and infrastructural asymmetries: the most advanced general-purpose AI models are developed and deployed by U.S.-based companies, many of which also dominate the digital platforms used to distribute creative content globally.[319] As generative AI outputs increasingly populate algorithmic feeds on dominant non-EU platforms, the EU risks deepening its structural dependency on external actors—undermining both media pluralism and cultural sovereignty. Existing instruments such as the Audiovisual Media Services Directive (AVMSD),[320] while valuable, were not designed to address the volume, velocity, and opacity of AI-generated content flows.[321]

---

316 See e.g. Ginsburg, J.C., & Budiardjo, L.A., Authors and Machines, 34 Berkeley Tech. L.J. 343, 445 (2019) (warning that granting copyright protection to authorless AI outputs risks distorting incentive structures and undermining the foundational coherence of copyright law).

317 See e.g. Pamela Samuelson, Generative AI Meets Copyright, cit., at 158–159 (highlighting the risk that generative AI may displace human authors in licensing and creative markets); Tim W. Dornis, The Training of Generative AI is Not Text and Data Mining cit. at 70-71 (analyzing the substitutive nature of generative outputs); Katherine Lee, et al., Talkin' 'Bout AI Generation: Copyright and the Generative-AI Supply Chain, cit., at 77–79 (noting how generative outputs may crowd out human-created content); Deven R. Desai & Mark Riedl, Between Copyright and Computer Science: The Law and Ethics of Generative AI, 22 Nw. J. Tech. & Intell. Prop. 55, 70–75 (2024) (exploring how commercial deployment of LLMs based on copyrighted training data may displace original works and destabilize creative markets).

318 See Tobias Holzmüller, personal communication, (May 2025).

319 See e.g. European Commission, Proposal for a Regulation on Contestable and Fair Markets in the Digital Sector (Digital Markets Act) COM(2020) 842 final (Dec. 15, 2020), Explanatory Memorandum, at 1–3 and Recitals (3)–(6) (highlighting the entrenched position of a few global gatekeepers, their control over content-distribution channels, and the structural dependencies that result).

320 See Directive (EU) 2018/1808, of the European Parliament and of the Council of 14 Nov. 2018 Amending Directive 2010/13 on the Coordination of Certain Provisions Laid Down by Law, Regulation or Administrative Action in Member States Concerning the Provision of Audiovisual Media Services (Audiovisual Media Services Directive) in View of Changing Market Realities, 2018 O.J. (L 303) 69-92.

321 See e.g. European Audiovisual Observatory, AI and the Audiovisual Sector: Navigating the Current Legal Landscape, IRIS, European Audiovisual Observatory, Strasbourg, 2024, at ch. 8 and 9 (highlighting the growing inadequacy of legacy

From an economic perspective, generative AI substantially alters established creative value chains.[322] By producing content rapidly, at scale, and often with quality comparable to that of human creators, generative AI introduces substitution effects that may erode creators' revenues and market standing. Economic theory suggests that as AI-generated content becomes increasingly abundant and cost-effective, consumer demand for human-authored works may decline, weakening incentives for original creation.[323] Moreover, the industrial scale of AI content production risks shifting bargaining power toward large technology platforms and intermediaries, contributing to market concentration and reduced diversity in cultural production.[324]

Legally, these market dynamics expose important gaps in fairness, attribution, and competition frameworks. The lack of clear rules on the status of AI-generated outputs and the role of human involvement in authorship creates uncertainty, complicates licensing and remuneration systems, and increases the risk of legal disputes. In the absence of targeted regulation, the existing framework risks undermining authors' rights and destabilising the creative economy.

Addressing these intertwined legal and economic challenges requires acknowledging that AI-generated outputs differ fundamentally from traditional creative processes in their market effects. A forward-looking regulatory approach must ensure fair competition, enable adequate remuneration for creators whose works underpin AI training, and guarantee transparency for consumers regarding the origin of content.[325] By integrating economic considerations into legal design, the EU can construct a regulatory architecture capable of sustaining a vibrant, diverse, and equitable creative ecosystem in the age of generative AI. In order to respond effectively, regulatory frameworks must move beyond input-side compliance and engage with the broader structural implications of AI-driven creative production. Section 4 sets out potential models to achieve this balance through legal clarity, fair remuneration, and institutional coordination.

frameworks such as the AVMSD in governing AI-driven distribution models, algorithmic personalisation, and media pluralism risks).

[322] See U.S. Copyright Office, Copyright and Artificial Intelligence, Part 3: Generative AI Training, cit. at 65. The Office identifies "market dilution" as a novel form of harm under the fourth fair use factor, noting that "the speed and scale at which AI systems generate content pose a serious risk of diluting markets for works of the same kind as in their training data." Even when outputs are not directly infringing, their stylistic imitation and volume may diminish the value of original works, raising broader economic concerns.

[323] See e.g. Ajay Agrawal et al., *Prediction Machines: The Simple Economics of Artificial Intelligence* (2018) at 37–39 (explaining how a drastic reduction in the cost of prediction leads to substitution effects, reshaping market dynamics and reducing demand for more expensive human inputs).

[324] See e.g. European Commission, Report of the High-Level Expert Group on the Impact of the Digital Transformation on EU Labour Markets, at 19–20, 44 (Eur. Comm'n, Apr. 2019), https://digital-strategy.ec.europa.eu/en/news/final-report-high-level-expert-group-impact-digital-transformation-eu-labour-markets; Andrei Hagiu and Julian Wright, Artificial intelligence and competition policy, International Journal of Industrial Organization 2025 (arguing that AI could lead to new types of gatekeepers).

[325] See e.g. Giuseppe Colangelo, A Competition Policy Analysis of Copyright Protection in Gen AI, Singapore Journal of Legal Studies, forthcoming (2025). Available at https://ssrn.com/abstract=5201510 (proposing to align copyright exception analysis with antitrust principles to assess the substitutive impact of GenAI outputs.).

3.4. Infringement and liability

Although AI-generated content is not protected under copyright law due to the absence of human authorship, it may still infringe the rights of existing rightsholders. Copyright infringement arises not from the identity of the creator but from the act of reproducing a protected work without authorisation. Accordingly, outputs generated by generative AI systems may be unlawful where they incorporate—directly or indirectly—elements of pre-existing copyrighted material used in the training process or otherwise accessed during generation.

A central question is where to draw the line between lawful inspiration and unlawful reproduction. Infringement is likely where an output includes a substantial part or recognisable fragment of a protected work. This could occur even if the output is combined by an AI system, with no intent to copy on the part of the user or provider. The situation is analogous to human infringement: if a person cuts and pastes a passage from a novel or a fragment of a copyrighted song, infringement arises regardless of intent—this same principle applies to outputs regurgitated by an AI model. Surveyed experts in a recent EU study expressed that even unrecognisable training use may infringe the moral right of integrity, especially when outputs mimic style or distort the author's reputation.[326]

There is a spectrum of possible scenarios. At one end is verbatim or near-verbatim reproduction—where a portion of a training work is reproduced almost identically, without significant modification. This phenomenon, while rare, has nonetheless been encountered in real-world contexts. For example, AI image models have occasionally reproduced images containing visible watermarks, suggesting direct storage and reproduction of training data.[327] Similarly, language models have been shown to output excerpts from books or documents contained in their training sets.[328] In such cases, the output is essentially an unauthorised copy of a protected work and would constitute clear infringement under EU copyright law. Even when outputs are probabilistic or described as "hallucinated," this does not preclude a finding of reproduction. When generative models return long or distinctive textual sequences that closely mirror protected material—especially under repeated prompting—the statistical likelihood of such outputs emerging without exposure to the original is extremely low. Empirical studies confirm that large models can and do memorize training data, supporting the inference that reproduction has occurred, even in the absence of a one-to-one match.[329]

---

[326] See European Commission, Study on Copyright and New Technologies: Copyright Data Management and Artificial Intelligence, cit, at 230.

[327] See e.g. Getty Images (US), Inc. v. Stability AI Ltd., [2023] EWHC (Ch) 3090 (UK High Court).

[328] One example is the New York Times v. OpenAI lawsuit, which highlights allegations that OpenAI's language models reproduced verbatim excerpts from paywalled articles, allegedly demonstrating the memorization and regurgitation of protected training data.

[329] See e.g. A. Feder Cooper and James Grimmelmann, The Files are in the Computer: On Copyright, Memorization, and Generative AI, 98 Chi.-Kent L. Rev. (forthcoming), at 48-49, available at SSRN: https://ssrn.com/abstract=4803118; Katherine Lee et al., Deduplicating Training Data Makes Language Models Better, in arXiv (2022) https://arxiv.org/abs/2107.06499; Peter Henderson et al., Foundation Models and Fair Use, 24 Journal of Machine Learning Research 1-79 (2023); Jooyoung Lee et al., Do Language Models Plagiarize? In Proceedings of the ACM WebConference 2023 3637–3647 (2023) https://doi.org/10.1145/3543507.3583199; Nicholas Carlini et al., Quantifying Memorization Across Neural Language Models, Eleventh Int'l Conf. on Learning Rep. (ICLR) (Mar. 6, 2023),

Beyond reproduction, some AI-generated outputs—particularly those disseminated through chatbots or automated search agents—may constitute a communication to the public under Article 3(2) of the InfoSoc Directive. According to the CJEU's doctrine in Svensson,[330] Reha Training,[331] and Infopaq, the act of making protected content available to a "new public" without authorisation may itself constitute infringement, even if the content was previously accessible. This criterion is met where the use of generative AI circumvents the original licensing environment and delivers content to a public that was not contemplated by the rightholder, such as users accessing paraphrased or summarised versions of protected works through AI interfaces, rather than via the licensed source (e.g. a press website). Importantly, the test is not limited to literal reproduction: it applies in technologically neutral terms and encompasses situations where AI-generated outputs act as functional substitutes for protected works, thereby interfering with the original's exploitation. For instance, when AI systems produce stylised artistic renderings or condensed news summaries that fulfil the same demand as the original work, the user receives the expressive value of the content without triggering access to the licensed source. In such cases, the delivery channel—not just the content—matters. This output-side risk underscores the dual exposure of generative AI systems: first at the training phase (via reproduction), and second at the delivery phase (via communication to the public), particularly when dissemination routes bypass or displace licensed access models.

More commonly, AI outputs may bear substantial similarity to training materials without being exact copies. Infringement in these cases depends on whether the generated content appropriates the protected expression of the earlier work, not merely its ideas, themes, or concepts. For example, if an AI-generated musical composition echoes the melody or harmonic structure of a copyrighted song, it could be deemed infringing. Determining substantial similarity is inherently contextual and often requires expert analysis, much like plagiarism assessments in human-authored works.

A more complex legal and normative debate surrounds AI-generated outputs that imitate the "style" of a particular creator. Style or technique per se is not protected by copyright, as it forms part of the unprotectable idea-expression dichotomy. Hence, an AI-generated painting "in the style of Van Gogh" or "in the manner of a living artist" is unlikely to infringe copyright, provided it does not copy specific expressive elements. However, some creators have voiced concerns that systematic imitation of style by AI tools erodes artistic identity and market value.[332] While these concerns are valid, they may fall outside the scope of copyright and enter the domain of unfair competition or the potential recognition of a new sui generis right in artistic style—an option not currently contemplated under EU law.

The allocation of liability for infringing outputs remains therefore a complex issue. At first glance, the user of the generative AI system bears primary responsibility. Where a user prompts the system to generate content that is likely to reproduce a protected character or work (e.g. "draw Mickey Mouse"),

https://arxiv.org/pdf/2202.07646; Jamie Hayes, et al., Measuring memorization through probabilistic discoverable extraction, in arXiv (2025) https://arxiv.org/pdf/2410.19482

[330] Case C-466/12, Svensson v. Retriever Sverige AB, ECLI:EU:C:2014:76 (Feb. 13, 2014).

[331] Case C-117/15, Reha Training Gesellschaft für Sport- und Unfallrehabilitation mbH v. Gesellschaft für musikalische Aufführungs- und mechanische Vervielfältigungsrechte (GEMA), ECLI:EU:C:2016:379.

[332] See Ahmed Elgammal, AI Is Blurring the Definition of Artist, 107 Am. Scientist, 18-21 (2019).

the resulting output could constitute an unauthorised derivative work. The user may be liable for its creation, use, or distribution, akin to a human who reproduces a protected image without permission.

Provider liability is more nuanced. Under existing law, technology providers are generally not liable for infringing acts committed by users, unless they have contributed to or facilitated the infringement knowingly.[333] However, if a model is developed or configured in a way that predictably outputs infringing content—e.g., because it memorises and reproduces protected works—then providers could be exposed to forms of indirect or secondary liability, analogous to contributory infringement in other jurisdictions, depending on the level of knowledge and control over infringing outputs. Moreover, if providers train their models on copyright-protected data without appropriate licences or opt-out mechanisms, they may already incur liability at the training stage. The AI Act introduces important obligations in this regard, requiring providers of general-purpose AI systems to implement "state-of-the-art" safeguards to prevent unlawful outputs. While these obligations are regulatory in nature and do not establish new bases of copyright liability, failure to implement effective safeguards may increase both legal and reputational risks.

As for the outputs themselves, they do not enjoy copyright protection and cannot be the subject of exclusive rights. However, if an output incorporates protected expression from a training work, any further reproduction, display or dissemination of that output could amount to copyright infringement of the original. This creates risk for downstream users—individuals or platforms—who may unwittingly share or commercialise infringing outputs. This situation points to the potential utility of automated detection tools—such as similarity checkers or AI-specific content ID systems—to flag outputs that are too close to known works. The development and deployment of such tools could form part of a wider risk mitigation strategy, supported by transparency requirements under the AI Act.

A number of legal proceedings have already emerged around these issues. In the United States, class-action lawsuits against Stability AI and other developers allege that generated images constitute derivative works of the training materials, infringing the rights of artists.[334] In the UK, Getty Images has filed suit against Stability AI, asserting that some outputs bear traces of Getty's watermarked images, suggesting unauthorised reproduction.[335] While these are early-stage proceedings and not binding on EU courts, they illustrate the legal tensions that are likely to emerge across jurisdictions. Within the EU, no definitive judgments have yet addressed output infringement, but the general principles of copyright law— particularly the standard that even a part of a work may be protected if it reflects the author's intellectual creation[336]—remain applicable. The main challenge lies in detection and evidence, rather than in the substance of the legal framework.

[333] Regulation 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and Amending Directive 2000/31/EC (Digital Services Act), 2022 O.J. (L 277) 1.

[334] Andersen v. Stability AI Ltd., No. 3:23-CV-00201 (N.D. Cal. Jan. 13, 2023).

[335] Getty Images (US), Inc. v. Stability AI Ltd., [2023] EWHC (Ch) 3090 (UK High Court).

[336] Infopaq Int'l A/S v. Danske Dagblades Forening, Case C-5/08, EU:C:2009:465, ¶¶ 37–39 (holding that even a part of a work—such as an extract of 11 words—is protected by copyright if it reflects the author's own intellectual creation).

As already mentioned, from a regulatory standpoint, the AI Act can play a complementary role. By requiring general-purpose AI model providers to implement safeguards, and by promoting transparency obligations, the Act helps address the upstream risk of infringement. Downstream, the Digital Services Act (DSA)[337] and the CDSM Directive (notably Article 17) may still apply when infringing AI-generated content is uploaded to online platforms, triggering notice-and-takedown obligations. However, these regimes were not designed with generative AI in mind, and new enforcement tools may be required to address the unique characteristics of AI-generated content.

Preventive measures are therefore crucial. These include technical filters embedded in AI models to prevent the reproduction of large verbatim passages or protected images, watermarking of AI-generated content to support traceability, and the creation of databases of known protected works to enable output comparison. Encouraging the development and adoption of such tools—possibly through industry codes of conduct or public-private partnerships—could help reduce infringement risks and build trust among creators and users alike. Looking ahead, emerging technical tools such as cryptographic watermarking or blockchain-based provenance tracking may improve the verifiability of dataset claims. While these are not yet mature enough for immediate deployment, the EU should support further R&D and standard-setting in this area.

Finally, it is important to clarify that the current EU framework provides a closed list of exceptions and does not recognise a general fair use defence. As a result, AI-generated outputs that include protected expression without a valid exception remain unlawful.

While AI-generated content may lack copyright protection itself, it remains subject to existing copyright constraints. Infringing outputs are unlawful, and liability may attach to users, providers, or platforms depending on the circumstances. The current legal framework appears adequate in substance, but its effective enforcement in the AI context requires technical support, regulatory coordination, and increased transparency across the AI value chain. This conclusion echoes the study's broader recommendation to avoid creating new copyright exceptions for AI-generated content and instead strengthen existing enforcement tools, safeguards, and transparency obligations to mitigate risks of unlawful reproduction.

[337] Regulation 2022/2065 of the European Parliament and of the Council of 19 Oct. 2022, on a Single Market for Digital Services and Amending Directive 2000/31/EC (Digital Services Act), O.J. (L 277) 1 (EU).

4. POLICY OPTIONS AND RECOMMENDATIONS

Generative AI blurs the boundaries of current copyright law. On the one hand, AI models ingest vast quantities of human-created content—often without authorization or compensation—to fuel their performance. On the other, they produce outputs that can mimic, substitute, or even surpass original works, yet lack a clear legal status. This dual disruption raises urgent questions about both the legitimacy of large-scale use of protected materials for training (input side), and the originality and ownership of AI-generated results (output side).

In navigating this terrain, we must strike a balance that neither hinders technological progress (and the benefits of AI for education, accessibility, innovation) nor undermines the foundations of the creative economy and the incentives for human authorship. As the European legal tradition consistently affirms, legal innovation must uphold the foundational principles of fairness, proportionality, and the integrity of authors' rights.

This concern becomes especially acute when considering the use of protected works in the training phase of AI systems. As the analysis in this study has demonstrated, while questions about the originality and potential infringement of AI-generated outputs are important, they are **comparatively less problematic than the unresolved and systemic legal uncertainties surrounding the ingestion of protected content during training**. This is where the most significant regulatory gaps—and risks—lie. While legal debate around AI training has gained traction, the broader economic disruptions caused by generative AI remain largely underexplored in legal discourse. These structural shifts—reshaping value chains, redistributing bargaining power, and altering revenue flows—are essential to address if we are to design a fair and future-oriented regulatory framework.

From the legal perspective, one of the central conceptual challenges is whether the mechanisms of machine learning can be meaningfully assimilated to human cognitive processes under copyright law. This question is pivotal. If machine "learning" were functionally equivalent to human study—reading a book, observing a painting, listening to a song—then the ingestion of protected content might not constitute an act of reproduction. However, if training entails large-scale copying, internal storage, and syntactic recombination of protected expression, then such use must be considered reproduction and require authorisation, or at least fall within clearly defined exceptions. While it is often suggested that AI systems "learn" in ways similar to humans—such as reading a book or studying a painting—this analogy is misleading from a legal perspective. Under EU copyright law, this study finds that such a comparison does not hold. When generative AI models are trained on protected content, they typically make copies and process the actual expressions found in those works. This goes beyond what is permitted under current legal exceptions for activities like research or analysis (see Section 2.1.2).

Unlike human authors, who understand ideas and express them in new ways, AI systems do not "understand" what they process. As philosopher Luciano Floridi puts it, AI acts without understanding—it follows statistical patterns rather than engaging with meaning. This difference matters legally.[338] A

[338] See Luciano Floridi, AI as Agency Without Intelligence: On ChatGPT, Large Language Models, and Other Generative Models, cit

person can learn from a work and restate the ideas in their own words without infringing copyright. But an AI system must copy and recombine parts of protected works to function. Even if the final output looks different from the training data, the act of ingesting and using protected content is still legally considered reproduction—and may require permission from rightsholders unless a clear legal exception applies.

In practice, the architecture of generative AI models—often described as black boxes—makes it nearly impossible to verify how content is used during training. Even developers may not fully understand how expressive elements are processed, stored, or transformed. This technical opacity compounds the legal and ethical concerns already discussed. Without transparency, it cannot be assumed that the system merely extracts neutral facts or patterns. On the contrary, there is a credible risk that substantial expressive elements are copied, internalised, and re-emitted through derivative outputs. The result is a process that closely resembles large-scale, unlicensed content reproduction.

In this context, **the precautionary principle**[339] becomes particularly relevant. If applied to generative AI, this principle may support a proactive regulatory stance. Legal uncertainty about the scope and effect of training practices should not delay the adoption of safeguards—especially when the risks include the displacement of creative labour, the dilution of economic rights, and the long-term erosion of Europe's cultural and knowledge ecosystems.

A useful analogy can be drawn with food safety regulation, where the precautionary principle is well established. Imagine a hamburger produced by mixing together traces of dozens of meats from unknown sources, with no indication of their origin, quality, or safety. Even if the final product appears edible and appealing, such opacity would be unacceptable in a sector that directly impacts human health. Strict regulations require traceability, disclosure of ingredients, and verifiable sourcing—to preserve public safety and consumer trust.

Yet a similar lack of transparency applies to the datasets used in generative AI, where the origin and legal status of content remain largely unknown. These systems ingest and process immense volumes of cultural and intellectual content—texts, images, audio—without clear provenance, permission, or oversight. The final outputs may appear innovative and valuable, but we have no visibility over what materials were used, how they were processed, or whether they included protected expression. **We do not know "what is** inside **the burger**."

Analogous to regulatory expectations in sectors such as food safety, **AI governance requires traceability and transparency** in order to ensure responsible data practice. There is no compelling reason to accept lower regulatory standards simply because the product is digital rather than edible. The raw material of generative AI is human creativity—arguably no less vital to the public interest than food—and it must be treated with equivalent care.

[339] On the precautionary principle, see Article 191(2) TFEU. See also Communication from the Commission on the Precautionary Principle, COM(2000) 1 final, 2 February 2000, which clarifies its broader application beyond environmental matters. As articulated by the European Commission, "recourse to the precautionary principle presupposes that potentially dangerous effects have been identified, and that scientific evaluation does not allow the risk to be determined with sufficient certainty."

A growing concern is the potentially unlawful ingestion of protected works without authorisation or transparency. This may amount to large-scale reproduction in breach of EU copyright law—particularly in the absence of a valid exception. Beyond legality, there is a deeper structural imbalance: AI developers derive immense value from copyrighted content without compensation or cost-sharing. This undermines sustainable incentives for creativity and accelerates the consolidation of power among dominant platforms. If unaddressed, this dynamic threatens economic fairness and the cultural integrity of Europe's creative ecosystem.

In this light, copyright law must be understood not as a barrier to innovation, but as a vehicle for ensuring that innovation remains ethically grounded and socially legitimate. Upholding the principles of stewardship and fairness is essential if AI development is to proceed within a framework that respects both fundamental rights and the public interest.

Accordingly, the policy recommendations that follow are structured around three guiding objectives:

- **Legal clarity:** refining the scope of permissible AI training practices and the status of AI-generated outputs.
- **Transparency and accountability:** enabling dataset traceability and auditability, while replacing the unworkable opt-out mechanism with a principled opt-in framework for AI training.
- **Fair** remuneration**:** establishing mechanisms to ensure that those whose works are used in training receive equitable compensation.

Collectively, these objectives operationalise the EU's precautionary principle: when large-scale, opaque data uses create systemic risks, regulators should front-load transparency, traceability and fair-value measures before harm materialises.

By embracing these principles, the European Union can guide generative AI development in a direction that aligns with its legal values of fairness, innovation, and cultural sustainability.

The following policy options are designed to translate the EU's core legal principles into concrete regulatory safeguards. They aim to ensure that the development and deployment of generative AI unfolds within a framework that respects authors' rights, prevents systemic imbalances, and promotes sustainable innovation across the Digital Single Market.

The following paragraph opens with an accountability test that will be used to evaluate each policy option.

4.0. Three-Pillar Accountability Test (orientation tool for Sections 4.1–4.6)

This section introduces an original policy evaluation tool: the **Three-Pillar Accountability Test**. It is designed to assess whether legal rules addressing generative AI uphold the core requirements of **transparency**, **fairness**, and **enforceability**. The same regulatory challenges that shaped earlier debates on digital platforms now apply with equal force to large-scale AI training and output markets. This framework identifies legal gaps that prevent the system from functioning accountably and guides the design of practical remedies to restore balance.

How to read this section.

The tables in this section apply a Three-Pillar Accountability Test to identify where AI-related copyright rules succeed or fall short. The test checks three basic conditions:

1) Transparency (epistemic) – Can creators see how their work is used?

2) Fairness (normative) – Are rights and revenues shared appropriately?

3) Enforcement (systemic) – Is there an EU-level body to uphold the rules?

If one of these elements is missing, the system remains unbalanced. Each table shows the current legal gap (left) and a concrete solution (right), so policymakers can quickly identify where legal reinforcement or reform is needed.

Table 6: The pillars at a glance

Pillar	Checks ...	Missing today	What a fix looks like
Epistemic *Who knows what?*	Can creators see whether, how and where their works are used?	Dataset summaries are unverified; opt-out tags vary by site.	• *Create* one EU-wide, machine-readable "do-not-train" tag • *Run* random audits of training corpora against that tag
Normative *Who sets the rules?*	Are rights and revenues fairly allocated?	Article 4 CDSM shifts the burden to creators; no pay-back for training use.	• *Introduce* a statutory collective licence → rightsholders fund
Systemic *Who polices compliance?*	Is there an enforcement body?	EUIPO lacks audit power; the AI Office lacks an IP brief; courts act case-by-case.	• *Create* an AI & Copyright Unit within the AI Office, in coordination with EUIPO and CMOs, to audit datasets and recommend enforcement actions.

Table 7: Why the pillars matter – a quick walk-through

AI-copyright stage	Where the gap bites	One real-world illustration
Training	Epistemic & Normative	**French publishers vs. OpenAI (2023):** newspapers learned of scraping only after code leaks; no verification roadmap exists.
Model deployment	Normative	Users can remix your graphic novel style without a licence; liability is pushed onto end-users via terms-of-service.
Output distribution	Epistemic & Systemic	Streaming sites host AI-generated songs without provenance labels; creators must sue individually.

Table 8: A "traffic-light" test for draft amendments

● = gap open ● = gap addressed through realistic EU mechanisms

Question to ask	● Red (gap still open)	● Green (gap addressed)
EP1. Can a creator discover whether her work sat in the training set?	Only a voluntary transparency report	Mandatory dataset log + spot audits
NP1. Does the rule guarantee income or veto power?	Opt-out only, no payments	Collective licence with revenue share
SP1. Is there an EU body that can order correction or suspension?	Pure court litigation	AI & Copyright Unit with coordinated audit oversight and escalation to enforcement bodies

➔ **Rule of thumb:** *If a proposal is green in all three rows, it passes the Three-Pillar Accountability Test*

How this helps the JURI Committee right now

• **Clarifies priorities:** highlights where transparency without audit (● EP-1) or remuneration without enforcement (● SP-1) leaves the core problem unsolved.

• **Filters options:** flags half-measures before legislative time is spent.

• **Future-proofs law:** any new AI practice – text-to-video, voice cloning, synthetic audio – can be screened with the same traffic-light grid.

The policy recommendations that follow are structured around a temporal logic: immediate interventions should focus on restoring legal clarity and enforcing existing rights under current law, while longer-term measures aim to recalibrate the legal framework through targeted reform.

4.1. Governance and enforcement: Fragmented responsibilities

The current institutional framework for managing the intersection of copyright and generative AI in the EU is fragmented, reactive, and not well-suited to the scale or complexity of the challenges ahead. Responsibilities are dispersed across national authorities, EU institutions, and enforcement bodies, resulting in regulatory gaps, duplication of efforts, and limited strategic coordination. This section outlines three mutually reinforcing proposals to address these governance deficiencies. The goal is to combine short-term expert input with longer-term institutional oversight and structured stakeholder dialogue—thereby creating a more coherent, resilient, and innovation-sensitive governance architecture for copyright in the age of AI.

A) Establish a permanent cross-sectoral governance platform

In order to support regulatory coherence and build long-term trust, the EU should establish a **permanent cross-sectoral platform for dialogue on AI and copyright**. This could take the form of:

– a dedicated **working group under the EU Observatory on IP Infringements**, expanded to address AI and creativity;

– or a new **multi-stakeholder forum** convened by the Commission or Parliament (e.g., under the auspices of the JURI Committee), bringing together rights holders, developers, platforms, regulators, and researchers.

Such a body should not be merely consultative. Its mandate could include monitoring emerging practices, proposing voluntary codes of conduct, contributing to soft law instruments, and advising on legislative updates. Structured dialogue of this nature is essential to keeping regulatory approaches current and ensuring that legal norms evolve in tandem with technological and economic developments. This proposal is consistent with the European Parliament's 2020 resolution, which highlighted the importance of cross-sectoral dialogue, open access for research, the development of technical standards for AI systems, and the need for human oversight and transparency in AI-assisted IPR enforcement.[340]

Building on this model of collaborative governance, a more operational mechanism may be required to address oversight, enforcement, and the economic implications of generative AI in the copyright domain. Rather than establishing a new institutional body—an approach that may entail significant legal, administrative, and budgetary complexity—this study proposes the creation of a dedicated **AI & Copyright Unit embedded within the EU AI Office**, and operating in close coordination with EUIPO and collective management organisations (CMOs). It is important to distinguish this proposal from the forthcoming EUIPO EUIPO Copyright Knowledge Centre, announced for launch in November 2025.[341] While the Knowledge Centre is expected to serve as a strategic hub for copyright-related resources, guidance, and stakeholder engagement, it is not intended to perform operational compliance or audit functions. By contrast, the **AI-Copyright Unit** proposed here would fulfil concrete governance tasks—such as dataset transparency verification, opt-out enforcement, providing technical advice on licensing models, and monitoring emerging practices at the intersection of copyright and AI—thus complementing, rather than duplicating, EUIPO's more knowledge-oriented role.

While the establishment of a fully independent **AI–Copyright Unit** could remain a valuable long-term option—especially if enforcement gaps persist or sectoral complexity grows—the immediate priority should be to consolidate copyright-related transparency and compliance tasks within an existing structure. This phased implementation would maximise feasibility and ensure alignment with the institutional logic of the AI Act, particularly Articles 64–66 on market surveillance and coordination.
The Unit's activities should remain grounded in legal due process and benefit from continuous input from stakeholders across sectors, including rightsholders, AI developers, civil society, CMOs, and Member State experts. An initial focus on soft law development, voluntary compliance mechanisms, and coordinated dataset audits could allow the Unit to demonstrate value, build trust, and inform future discussions on whether a standalone enforcement entity—such as a Board—would later be warranted.

[340] See European Parliament resolution of 20 October 2020 on intellectual property rights for the development of artificial intelligence technologies, 2020/2015(INI), 2021 O.J. (C 404) 129, at §§10–11 and §16.

[341] See EUIPO, Strategic Plan 2030. Available at https://www.euipo.europa.eu/en/about-us/governance/strategic-plan

The Unit would thus serve as a key institutional interface between copyright law and AI governance, ensuring that legal obligations under the AI Act are interpreted and implemented in a manner that respects EU copyright principles and stakeholder interests.

B) Establish a JURI Working Group on AI and Copyright

At the parliamentary level, a key structural measure would be the establishment of a dedicated Working Group within the Committee on Legal Affairs (JURI) to ensure institutional continuity and strengthen the European Parliament's long-term engagement with the evolving interface between copyright and artificial intelligence. This Working Group would serve as **an internal parliamentary structure** tasked with supporting horizontal coordination, democratic oversight, and legislative follow-up. The Working Group would also enhance the Parliament's capacity for independent legislative foresight in this domain, without relying on external advisory processes. It could provide a **forum for structured dialogue** with other committees (such as IMCO, CULT, or ITRE), ensure that stakeholder perspectives continue to inform the parliamentary debate, and help identify gaps in the existing legal framework. The Working Group could also monitor forthcoming jurisprudential developments—such as the ruling in Case C-250/25—and prepare the ground for future regulatory initiatives on copyright and generative AI. By embedding this topic within the Parliament's internal structures, the Working Group would reinforce **systemic accountability and cross-sector coherence,** ensuring that the EU's copyright framework remains responsive to technological innovation and aligned with fundamental rights.

C) Launch a time-limited High-Level Expert Group on AI Training & Copyright

In parallel with the establishment of longer-term governance structures, the European Commission—acting upon a strong resolution of the European Parliament and in collaboration with the EU AI Office—should convene **a High-Level Expert Group (HLEG) on AI Training & Copyright by Q4 2025**. The interaction between copyright and generative AI training is a profoundly cross-cutting issue—touching on legal, technical, economic, and cultural dimensions—and **no single actor** (DG CONNECT, DG JUST, EUIPO, or the AI Office) **currently holds the full institutional or disciplinary picture**. **A time-limited HLEG provides the most structurally appropriate and inclusive mechanism to consolidate expertise, build consensus, and generate technically actionable outputs**. This approach is well-grounded in precedent: past groups such as the AI-HLEG (2018) and the Article 17 DSM Stakeholder Dialogue (2019) played a decisive normative role, especially where their conclusions informed delegated acts or operational codes of practice.

Importantly, the proposal is designed to complement—not delay—ongoing policy action. While the GPAI Code of Practice under Article 53 of the AI Act is advancing, it remains necessarily general in scope. A copyright-specific HLEG would provide domain-focused input that could directly support the finalisation of sectoral annexes and inform the Commission's 2026 review of Articles 3 and 4 of the CDSM Directive. The group's mandate should be narrowly scoped and time-limited (six months), delivering concrete technical and legal recommendations by mid-2026. In the current legislative climate, where Member States have signalled reluctance to reopen the CDSM Directive in the short term, and where non-binding guidance risks limited uptake without political consensus, the HLEG represents a pragmatic and timely mechanism to underpin balanced, forward-looking copyright reform in the age of AI.

1. Rationale and Purpose

To address the legal uncertainty surrounding the use of copyrighted content in AI training and to support legislative and regulatory implementation, this HLEG would focus on clarifying the opt-out mechanism under Article 4, supporting auditability of training datasets, and exploring feasible remuneration models for rightsholders. Its work would contribute directly to the Commission's 2026 review of the CDSM Directive and to the evolving governance of generative AI in the EU.

2. Political Space

There is strong political momentum for such an initiative. Member States have called for clarification of the opt-out regime and further standardisation of dataset transparency. Simultaneously, stakeholders across the creative and technology sectors demand clear, fair, and technically feasible rules. Parliament, by supporting this proposal, can position itself as a facilitator of consensus and a proactive actor in bridging rights protection with innovation.

3. Timing and Duration

The HLEG should be launched by Q4 2025 with a six/nine-month mandate, and deliver its conclusions no later than 30 September 2026, to ensure alignment with:

- The Commission's evaluation of Articles 3–4 of the CDSM Directive.

- The finalisation or implementation phase of the General-Purpose AI Code of Practice under the AI Act.

4. Scope of Work

The group should deliver:

- A standardised, machine-readable opt-out syntax under Article 4 CDSM (e.g., IPTC metadata, C2PA, robots.txt).

- A structured audit template for dataset summaries required by Article 53 of the AI Act, including minimum information requirements and verifiability standards.

- A detailed assessment of policy options for a pilot statutory remuneration mechanism or extended collective licensing model for AI training uses.

5. Complementarity with Existing Processes

This proposal is not in conflict with the work of the Chairs and Vice-Chairs currently drafting the GPAI Code of Practice. Rather, it provides copyright-specific, technical input that complements the broader scope of the Code. Its outputs can directly feed into the Code's final sectoral annexes and the Parliament's future legislative agenda.

6. How to Keep It from Slowing Things Down

To avoid duplication and delays:

- The group's scope and deadlines should be clearly predefined.

- It should be embedded within or report to the EU AI Office in coordination with the Commission, rather than operate as a standalone entity.

– Its work should run in parallel with the implementation of the AI Act and other enforcement mechanisms.

7. Parliament's Toolbox

While Parliament cannot create the group directly, it can request it formally through a resolution, own-initiative report, or in the context of the Article 225 TFEU. This mirrors past practice:

– The AI-HLEG (2018), following the Parliament's robotics resolution (2015/2103(INL))

– The Data Act Expert Group (2022) (2019/2180(INI))

A similar call could be incorporated into a legislative resolution on the implementation of the AI Act or the CDSM review.

Table 9: Three-Pillar Check

Pillar	Status	Why?
Epistemic	🟡	The proposed AI & Copyright Unit would improve transparency by verifying dataset disclosures and opt-outs, but lacks statutory powers to compel data access or enforce uniform standards across platforms.
Normative	🟡	Encourages fairer outcomes via structured dialogue and legislative foresight (e.g., JURI Working Group), but does not yet establish binding rights or remuneration mechanisms.
Systemic	🟡	Lays the groundwork for stronger enforcement by introducing institutional coordination tools (Unit, Working Group, HLEG). While no binding mandate exists yet, the phased approach allows for future escalation into an enforcement-capable structure.

Note: These proposals provide the institutional foundation for future reform. To meet full three-pillar accountability, they must be complemented by legal instruments that introduce enforceable rights and mechanisms (see §4.2–4.4).

4.2. Improve implementations of TDM exceptions

A) Why interim measures are still needed

These short-term actions **do not legitimise** the use of Article 4 for generative-AI training. Rather, they are intended to **reduce legal fragmentation** in the interim—**only for bona fide text-and-data mining** (TDM) where acts of reproduction are permitted because their sole aim is analytical, not synthetic or expressive (see § 4.2). Building on the analysis in Section 2, the measures below aim to provide legal clarity for strictly analytical uses, **without prejudging** the future legal status of Article 4 in the generative-AI context.

While Article 4 may **appear, at first glance**, to offer a lawful basis for generative AI training, a closer reading—of both the Directive's wording and the functional realities of generative AI—reveals that this interpretation is **legally strained, technologically unfounded, and normatively troubling**. The interim measures proposed below are designed to harmonise genuine, non-expressive TDM. They do not

extend Article 4's application, nor do they endorse its use for generative-AI training. Stakeholders relying on Article 4 to justify such training do so at their own legal risk until specific legislation is adopted.

Article 4 was introduced to support **data-driven innovation** by allowing certain users—under defined conditions—to carry out text and data mining without prior authorisation. It was intended to enable the extraction of patterns, trends, or factual correlations from large datasets, particularly in areas such as scientific research and data analysis. Its scope **was deliberately narrow**, to preserve the balance between enabling innovation and protecting the rights of authors and creators. This intention is clear from Article 2(2), which defines TDM as an "automated analytical technique aimed at analysing text and data… to generate information," and from Recital 8, which confirms that TDM is meant to derive **knowledge**, not creative expression, from the analysis of large volumes of data.

B) Clarify "information" vs "expression" under Article 4

In order to remove doubt, future guidance must state that Article 4 covers only analytical uses that **do not** internalise expressive form. Generative AI systems diverge fundamentally from this intended use. These models do not merely extract semantic content or identify correlations among facts. Rather, they absorb, encode, and recombine stylistic, structural, and expressive features of the works on which they are trained. This process enables the generation of outputs that can closely mimic the form and tone of the original works, blurring the line between inspiration and reproduction. Technically, this involves the transformation of input works into latent vector representations that preserve syntactic and stylistic information. Legally, this brings the process far closer to acts of reproduction than to acts of analysis.

This distinction is not semantic—it is foundational. Copyright law protects the expression of ideas, not the ideas themselves. While TDM in the narrow sense might target the latter, generative AI training targets the former. The reproduction of expressive form, whether literal or latent, engages the author's exclusive rights and cannot be equated with information extraction as contemplated by Article 4. As legal scholars have shown,[342] models trained on protected works are not just "analysing" data—they are "digesting" it and using it to recompose new outputs in the same expressive register. This is qualitatively and quantitatively different from the kind of pattern discovery the TDM exception was intended to permit.

C) Standardise opt-out and lawful-access conditions

A single, EU-wide rulebook is needed to make copyright reservations *technically visible* and legally enforceable during large-scale scraping:[343]

[342] See supra note 125.

[343] See EUIPO, *The Development of Generative Artificial Intelligence from a Copyright Perspective* (May 2025), cit. at 228–234 (See EUIPO, *The Development of Generative Artificial Intelligence from a Copyright Perspective* (May 2025), cit. at 228–234 (noting the lack of harmonised standards for opt-out tools, the fragmentation across protocols such as robots.txt, IPTC metadata, and C2PA, and the resulting ambiguity for developers and rightsholders alike). This ambiguity—though not directly questioned by the EUIPO—reinforces the concern, raised in this study, about the overall suitability of opt-out-based systems as a reliable compliance mechanism for AI training at scale.

Table 10: Standardise opt-out and lawful-access conditions

Layer	What it does	Who must comply
1. Machine-readable signal	Adopt a harmonised technical standard—e.g., an embedded "no-AI" field in IPTC/EXIF metadata and an upgraded robots.txt 2.0 protocol—so that any work or webpage can broadcast an opt-out in a format no crawler can ignore.	All hosting platforms, CMS vendors and dataset brokers.
2. Legal trigger	Require that developers (and dataset suppliers) invoking Article 4 demonstrate: (i) lawful access (licensed, subscribed, or public domain), and (ii) full compliance with any machine-readable opt-out signals. Failure moves the use outside the TDM exception and triggers ordinary infringement remedies	All entities conducting bona fide analytical TDM (e.g. universities, medical researchers, data-analytics firms).
3. Transparency hook	Require the same actors to file a brief Article 53 AI Act report listing content source categories and certifying compliance with the opt-out standard.	AI developers (or their EU representative).

This layered model turns today's voluntary opt-out signals into enforceable compliance mechanisms and clarifies the meaning of "lawful access." A harmonised signal lowers integration costs for developers, while a robust legal backstop reassures rightsholders that opt-outs will be respected.

Note — This standard is a compliance tool for lawful TDM only. It neither expands Article 4's scope nor legitimises the ingestion of protected content for generative AI training. While a standardised opt-out framework is often presented as a pragmatic compromise between innovation and copyright protection, this study adopts a more cautious stance. Fragmentation across technical standards, the rapid evolution of crawling technologies, and the uneven ability of creators to implement reservations—particularly for scraped or rehosted content—suggest that opt-out systems may not provide a reliable compliance baseline at scale. These concerns reinforce the need to explore more robust legal safeguards for generative AI.

D) Restore the Opt-In Principle: Reject Article 4 as a Legal Basis for GenAI Training

This subsection explains why ex-ante transparency schemes cannot cure the fundamental misfit between Article 4 CDSM and generative-AI training.

At its foundation, EU copyright law is an exclusive-rights regime: protection arises automatically for any work that meets the originality threshold—defined by the CJEU as the author's own intellectual creation. Any subsequent use—especially when it is large-scale, commercial or expressive—requires prior authorisation, unless a strictly limited exception applies. This principle, enshrined from the Berne Convention through the InfoSoc Directive, safeguards authorial control, proportionality and legal certainty.

Layering formal transparency on top of Article 4 would entrench the opposite logic. The opt-out model presumes that copying is lawful unless authors embed machine-readable reservations (robots.txt, IPTC, C2PA, etc.). That inversion of the burden effectively treats silence as consent. It would be akin to assuming that the contents of a book are freely reproducible unless the author prints "no copying" on

every page—undermining the very structure of exclusive rights. Introducing a presumption of lawfulness for the systematic extraction of protected content—absent an explicit opt-out—effectively converts the exclusive right into a default licensing regime. This approach mirrors the logic and operational features of open licensing models, yet lacks the element of consent that underpins such frameworks. As a result, it distorts the fundamental nature of copyright as a proprietary right and repositions it as a permissive instrument, oriented toward uses and objectives that diverge from its original normative rationale.

As discussed in section 2.1.3, opt-out tools do not prevent downstream reuse once content is stripped of metadata, rehosted, or transformed into screenshots, soundbites or synthetic data. Control lost at the training stage is irrecoverable, and transparency alone cannot compensate for the absence of initial consent. Empirical studies also show that small and independent creators bear a disproportionate administrative burden, while actual uptake of opt-outs remains minimal.[344] Furthermore, invoking scale and technical necessity as justifications risks establishing a precedent whereby technological constraints dictate legal rights. The unqualified acceptance of such logic would permit the circumvention of any right deemed "inconvenient" at scale, setting a problematic precedent for future regulatory decisions. A further example of this structural erosion—discussed in Section 2.1.3(a)—is the emerging practice of 'data laundering,' whereby datasets compiled under Article 3 for scientific research are subsequently reused in commercial AI training under Article 4.[345] This practice circumvents the intended limits of both provisions, allowing effectively commercial uses to benefit from a research-based exception. Reinstating an opt-in default would help restore the normative boundary between scientific and commercial text and data mining, and reassert the role of consent as a cornerstone of copyright governance.

Traditional exceptions—quotation, parody, private copying—permit narrowly delimited acts, each confined by purpose, scope and the three-step-test. Although exceptions and limitations are integral to copyright systems, they are typically structured with narrow scope and clear public-interest justifications. Article 3 of the CDSM Directive reflects this tradition by limiting TDM to scientific research. Article 4, by contrast, permits the wholesale reproduction of entire works for machine ingestion unless the author has effectively invoked an opt-out—thereby potentially altering the economic and moral equilibrium of copyright protection. Even arguing that the opt-out model is necessary to facilitate AI development, reduce transaction costs, and support Europe's digital competitiveness, the mechanisms chosen must remain consistent with legal principles. Altering the default licensing regime may increase the enforcement burden on individual creators and raise legal uncertainty in the absence of harmonised safeguards. There are less distortive options—such as collective licensing schemes, fair remuneration frameworks, or the creation of mandatory registries for dataset curators—that could achieve the same goals while respecting the structural integrity of copyright law.

[344] See supra §2.1.3

[345] See EUIPO, The Development of Generative Artificial Intelligence from a Copyright Perspective (May 2025), cit. at 117 (noting the risk of dataset reuse between Articles 3 and 4 CDSM, referred to as "data laundering").

This reconfiguration of the legal baseline is not merely a technical evolution. It has systemic implications, as it redefines the function of copyright from a system of individual control to one of default access—undermining the enforceability and clarity of the right itself. The opt-out regime under Article 4 represents not a mere exception, but a paradigmatic shift in the nature of copyright protection. It subordinates the exclusivity of rights to a presumed utility for innovation, enforced through mechanisms that structurally favour large-scale, well-resourced users over individual creators. This approach lacks both normative proportionality and practical safeguards.

Finally, the argument that Article 4 merely extends the list of traditional copyright exceptions downplays its structural novelty. Most exceptions operate unconditionally and do not require technical intervention by the rightsholder. Article 4(3), by contrast, conditions the exercise of rights on technological readiness—thereby creating an exclusionary effect against creators with fewer resources. Against this backdrop, the study recommends that the EU:

1) Reaffirm that training generative-AI systems on protected content requires prior, opt-in authorisation;

2) Support EU-wide licensing frameworks and rights-management systems based on affirmative consent;

3) Clarify legislatively that Article 4 was never intended to, and does not, extend to generative-AI training.

Restoring opt-in primacy is essential if EU copyright is to remain doctrinally coherent, technologically relevant and normatively sound in the era of generative AI.

This recommendation to restore the opt-in principle must also be understood as part of a phased regulatory strategy. In the short term, reaffirming that training on protected content requires prior authorisation is essential to halt unlicensed exploitation and re-establish a credible enforcement baseline. However, **the long-term viability of individualized licensing across billions of works is limited**. As discussed in Section 4.2, **a statutory licensing scheme or collective remuneration mechanism may ultimately provide a more scalable and equitable model**. These solutions would obviate the need for granular permissions while still ensuring that creators are fairly compensated. Therefore, restoring opt-in primacy should be seen not as a permanent end-state, but as a necessary transitional measure to preserve legal coherence while more systemic reforms are developed. This brings us to the question of feasibility. Importantly, the feasibility of structured prior authorisation mechanisms should not be underestimated. Prior consent remains the normative baseline of copyright law: the use of protected works is prohibited by default unless authorised in advance or clearly permitted under a limited exception. Just as social media platforms cannot function without users' prior consent—secured through non-exclusive licences—the same principle should apply to generative AI developers whose systems are built on the ingestion of protected content.

The argument that generative AI systems are "too complex" or "too large" to implement meaningful authorisation overlooks the fact that other large-scale digital infrastructures have already embedded licensing frameworks at scale. Rather than treating generative AI as a legal anomaly, the EU should

recognise it as a content-dependent infrastructure and apply the same logic of structured consent and transparent licensing that governs other platforms.

This comparison further illustrates that an opt-in model need not entail individual negotiations or high transaction costs. Just as platforms secure licences as a condition of use, AI developers could adopt system-level mechanisms to obtain prior authorisation before incorporating protected content into training datasets. Such an approach aligns fully with the principle of prior consent while remaining scalable and compatible with digital infrastructure realities.

Comparable licensing models already operate across the digital ecosystem. Audio-visual streaming services, for example, routinely rely on blanket licences or statutory schemes to ensure large-scale content access while preserving authorial control and remuneration. Moreover, several AI developers—including OpenAI —have begun to negotiate licensing arrangements with publishers, news organisations, and image banks.[346] These developments demonstrate that scalable licensing for high-volume content use is not only conceptually viable, but already emerging in practice. The notion that structured licensing would be unworkable in the AI context therefore lacks empirical support. Opt-in frameworks—particularly those based on tiered obligations and machine-readable permissions—can accommodate a range of actors, including open-source and non-profit projects. In this light, the claim that prior authorisation would impede innovation appears less a legal inevitability than a strategic policy choice—one that merits urgent reconsideration.

The following Box addresses the most common objections to an opt-in model—and explains why they do not hold.

Objection 1: Platforms deal with user-generated content, not third-party works.

This is **formally correct**, but it misses the core point: the issue is not who uploads the content, but whether **structural authorisation** exists. Social media platforms require users' prior consent via standard licensing terms before content is hosted or used. Likewise, AI developers can embed **scalable prior authorisation mechanisms**—especially when systematically ingesting large volumes of third-party content.

➔ **distinction between user-generated and third-party content:** It is true that AI systems ingest content originating from third parties rather than from direct uploaders. However, this reinforces—not weakens—the case for structured authorisation. If anything, the lack of a direct user relationship increases the legal and ethical responsibility of AI developers to establish robust consent mechanisms. Complexity in sourcing should not be used as a justification to bypass core principles of copyright law; instead, it calls for improved traceability, metadata standards, and licensing channels—solutions that already exist in other sectors managing distributed rights ownership.

Objection 2: The volume of data used for AI training makes licensing unmanageable.

[346] See supra note 137. See also EUIPO, The Development of Generative Artificial Intelligence from a Copyright Perspective (May 2025), cit. at 13.

This is an argument against **fragmented micro-licensing**, not against an opt-in model per se. Scalable solutions already exist—including **collective licensing**, **extended collective licensing (ECL)**, **central registries**, and **sectoral licensing hubs**—which demonstrate that opt-in frameworks are not inherently unworkable. The challenge is **infrastructural**, not legal.

➔ **On the feasibility of licensing at scale:** The current absence of comprehensive licensing infrastructure across all content types is a logistical challenge, but not a structural barrier to an opt-in system. Copyright history shows that rights management frameworks evolve in response to technological needs—whether through collective management organisations, extended licensing, or statutory interventions. Just as similar mechanisms were established to support broadcasting and streaming, legal and policy support can now facilitate their development for generative AI. Delaying action due to infrastructural inertia risks entrenching an unsustainable and inequitable system of de facto appropriation.

Objection 3: AI training repurposes existing content for a new function and therefore doesn't compete with original works.

This claim is both disputed and context-dependent. In many sectors—such as illustration, journalism, and stock photography—generative AI directly undermines existing markets. Moreover, under EU law, a use may infringe the reproduction right even if it serves a new purpose or adds value; the decisive criterion remains whether the use involves unauthorised reproduction of protected expression.

➔ **On the "repurposed" nature of AI training:** The claim that generative AI merely extracts general patterns or that its outputs are non-competing overlooks how models reproduce expressive elements of protected works, including structure, style, and creative syntax. Even if the output is not identical, the training process itself typically involves the reproduction of substantial portions of protected material—engaging the reproduction right under EU law. These developments have tangible substitution effects, calling into question the balance between access and incentives at the heart of copyright law. Unlicensed copying remains unlawful, particularly where viable alternatives such as structured licensing schemes exist.

Objection 4: Social media platforms aren't using content to train models or generate outputs.

Precisely—and this underscores the need for stronger safeguards in the AI context. Social platforms primarily **host** user content, whereas AI developers **process and reconfigure** protected works to generate new outputs that often replicate **style**, **structure**, or **substance**. This **qualitatively different use** intensifies the legal and ethical need for **prior authorisation.**

➔ **On the alleged qualitative difference between hosting and generation:** The fact that social platforms do not process content to generate new outputs actually underscores the legal asymmetry. AI systems do more than host—they analyse, model, and reconstitute protected expression, often in ways that are non-transparent and potentially prejudicial to authors' economic and reputational interests. If hosting requires licensing or platform terms, then a fortiori, the deeper transformative use carried out by AI should require at least equivalent safeguards. The higher the functional intensity of the use, the stronger the justification needed under copyright principles.

Objection 5: Opt-in systems would stifle European AI innovation and competitiveness.

This concern reflects broader fears about regulatory "**overreach**," but it **overstates the friction** of licensing while **underestimating the risks of unchecked appropriation**. Prior authorisation is not an obstacle to innovation—it is a **governance mechanism** that ensures **fair value distribution** and **market transparency**. Moreover, **licensing models are already emerging in practice** (e.g. OpenAI-AP partnerships), suggesting that consent-based innovation is viable. A well-designed opt-in model would **support legal certainty and sustainable development**, rather than hinder it.

➔ **On competitiveness and innovation risks:** Concerns about stifling European innovation are legitimate but should not be overstated. Prior authorisation is not a ban—it is a governance mechanism. Scalable, interoperable licensing systems—combined with tiered obligations based on scale and purpose—can accommodate both commercial innovation and the protection of authors' rights. Furthermore, a permissive regime that undermines European creators may paradoxically weaken Europe's long-term digital competitiveness by disincentivising quality content production and favouring data-rich incumbents. Aligning innovation with accountability is not only possible—it is essential for a balanced digital economy.

E) Practical Policy Recommendations

In order to enhance the implementation of Article 4 in a manner that respects the boundaries of copyright law while promoting legal certainty and convergence, the following actions are recommended:

1. Clarify the boundaries between "information" and "expression" under Article 4 through interpretative guidance or soft law instruments issued by the European Commission. These instruments should reaffirm that the TDM exception applies only to the extraction of semantic content for informational purposes, and does not extend to training processes that involve internalising and reproducing expressive elements of protected works.

2. Promote harmonisation among Member States by encouraging consistent approaches to the implementation of Article 4—limited to bona-fide, non-expressive TDM—particularly for lawful-access requirements, opt-out syntax, and machine-readable reservations. This harmonisation is an interim measure and will sunset once the Union adopts an opt-in framework for generative-AI training.

3. Pending formal clarification, any developer that *claims* Article 4 for AI-training must publish a comprehensive, ex-ante disclosure (datasets, legal theory relied on, opt-out screening). This obligation is purely defensive and does not prejudice subsequent infringement findings or the shift to an opt-in regime.

4. Apply the three-step test (InfoSoc 5(5)) rigorously: the exception cannot apply where large-scale or expressive ingestion prejudices normal exploitation. In practice, most generative-AI training will fail this test, reinforcing the need for an opt-in licensing route (see 2.1.2).

5. Clarify the relationship between the CDSM Directive and the AI Act by making explicit—possibly through a joint interpretative statement or delegated act—that Article 53(1)(c) of the AI Act does not

expand the material scope of Article 4, but merely affirms that developers must comply with copyright law as it stands.

These five actions are short-term safeguards. They neither endorse Article 4 for generative-AI training nor dilute the ultimate move to an opt-in scheme (see § 4.1-f and § 4.2).

F) Transition to a purpose-built solution

While targeted improvements to the TDM exception can reduce fragmentation in the short term, they cannot substitute for a coherent, opt-in legal framework capable of addressing the specificities of generative-AI training. **Article 4 of the CDSM Directive was never designed for the ingestion and recombination of expressive content at scale.** Attempts to retrofit this provision to accommodate generative AI risk undermining the protective logic of copyright, distorting the exception's original purpose, and eroding the incentive structures that support European creators.

For these reasons, the recommendations set out above should be seen as **interim, corrective measures**—necessary to contain the misapplication of existing law, but insufficient to govern the future of AI and creativity. The next section therefore turns to the more sustainable solution: the establishment of a dedicated EU-level exception or licensing mechanism for generative AI training, designed to balance innovation with authorial rights, economic fairness, and cultural sustainability.

Table 11: Three-Pillar Check

Pillar	Status	Why?
Epistemic	🟡	Introduces ex-ante disclosure obligations under Article 53 AI Act, improving transparency, but lacks enforceable audit mechanisms for compliance verification.
Normative	🟡	Clarifies the distinction between "information" and "expression" and reinforces the limits of the three-step test; however, no remuneration or consent mechanism is provided.
Systemic	🔴	Relies solely on court interpretation without establishing a dedicated enforcement body or transitional framework toward the proposed opt-in model (see §4.2).

Note: This measure stabilises the legal baseline in the short term but must be paired with structural reforms in 4.2 (remuneration) and 4.4 (traceability and safeguards) to achieve full three-pillar compliance.

4.3. Possible mechanisms for remuneration

A) Concept and legal rationale

In response to growing tensions between rightsholders and AI developers over the use of protected works for training generative AI models, this section outlines a precaution-driven, legally grounded and operationally feasible remuneration model that strikes a balance between innovation and creators'

rights. Unlike existing levy-based[347] or fundamental rights-based[348] proposals, this approach is structurally inspired by established EU mechanisms such as the cable retransmission regime, the artist's resale right, and Article 18 of the DSM Directive.

We propose the introduction of **a new EU-level statutory exception** to copyright for the specific purpose of training generative AI systems. This would be coupled with **an unwaivable right to equitable remuneration** for authors and rightsholders whose works are used in such training. This model reflects the reality that individual licensing is unworkable at the scale and speed of AI training, while ensuring creators are not excluded from value chains driven by data. This long-term model does not undermine the short-term necessity to reaffirm the opt-in principle. Rather, it acknowledges that the existing framework lacks the structural capacity to support large-scale compliance in the context of generative AI. More fundamentally, the proposed remuneration mechanism responds not only to fairness concerns but to a systemic market failure: human creators are being structurally excluded from value chains due to the industrial scale, speed, and substitutive effect of generative AI outputs. In such a context, even collective opt-ins or voluntary schemes are insufficient to rebalance negotiating power. A statutory exception coupled with an unwaivable remuneration right addresses this asymmetry and restores minimum economic agency to authors whose works underpin the system.[349]

Here, it is also important to note that these risks are not limited to input-side copying; they extend to the output-side market impact of generative AI. In this regard, the theory of market dilution—advanced by the U.S. Copyright Office in its recent report —raises novel concerns that may have direct relevance under the EU's proportionality principle and the three-step test (see 2.1.2).[350] When AI-generated content floods the market and stylistically imitates protected works, it may not amount to direct infringement but can nonetheless undermine the normal exploitation of the work. This may cause the exception to fail the third step of the Berne/TRIPS test and trigger the need for compensatory mechanisms under Article 18 of the CDSM Directive.

B) Why It Works:

Unlike private copying levies, which are premised on user-based consumption and often misaligned with AI training logic, this proposal draws inspiration from:

- Cable retransmission (Art. 9, Satellite and Cable Directive 93/83/EEC): where reuse occurs without individual prior consent but remuneration is guaranteed through collective systems.

---

[347] See Senftleben, Martin, Generative AI and Author Remuneration. 54 IIC - International Review of Intellectual Property and Competition Law, 1535–1560 (2023).

[348] See Geiger Christophe and Iaia Vincenzo, The forgotten creator: towards a statutory remuneration right for machine learning of generative AI. 52 Computer Law & Security Review 1-9 (2024).

[349] See Section 2.4 of this study (noting the bargaining asymmetry between authors and AI developers).

[350] See U.S. Copyright Office, Copyright and Artificial Intelligence, Part 3: Generative AI Training, cit. at 65. The report identifies "market dilution" as a novel harm, noting that "the speed and scale at which AI systems generate content pose a serious risk of diluting markets for works of the same kind as in their training data." Even without direct copying, such imitation and market saturation may affect the legitimate interests of authors and fall outside the scope of permissible limitations under international copyright norms.

– Artist's resale right (Directive 2001/84/EC): providing a proportional and non-waivable share of downstream value.

– Art. 18 DSM Directive: ensuring that creators receive appropriate and proportionate remuneration even in complex contractual chains.

C) Core Structure of the Model

The proposed scheme is built around three interdependent legal and operational components. Together, they enable a workable balance between legal access for AI developers and fair compensation for rightsholders. These components are summarised below:

Statutory Exception: A new provision in the EU copyright framework would permit training generative AI models using protected works, bypassing the need for individual authorization.

Remuneration Right: This use would trigger a mandatory, unwaivable right to equitable remuneration for the use of works as training inputs.

Collective Management: Remuneration would be collected and distributed by sector-specific collective management organisations (CMOs), in line with existing practices in music, audio-visual and visual arts.

D) Role of Collective Management Organisations (CMOs)

Collective Management Organisations (CMOs) are the practical hinge of the proposed remuneration scheme. Because they already collect and distribute royalties in music, audio-visual and visual-arts markets, they have the registers, matching engines and audit routines needed to handle AI-training payments at scale. CMO involvement also gives individual creators—especially smaller ones—a democratic, transparent channel to challenge allocations and track income. To make the system work EU-wide, the Commission should: (i) promote cross-sector data standards (hash+metadata), (ii) fund a single claims / opt-out portal, and (iii) facilitate reciprocal agreements so that a creator registered with one CMO is covered everywhere.

Several licensing levers that CMOs already use in other contexts can be repurposed for AI-training data. The table that follows lines up four realistic configurations, from the lightest voluntary option to the heaviest statutory back-stop.

However, it is important to acknowledge that not all creative sectors are currently represented by CMOs or have developed collective licensing infrastructures. While CMOs offer an efficient channel for remuneration in fields like music and visual arts, other sectors—such as software, academic publishing, or emerging digital formats—may require alternative governance models. In order to ensure inclusiveness and effectiveness, the proposed scheme should incorporate fallback mechanisms (e.g., through sector-specific funds, national registries, or EU-backed distribution bodies) in cases where CMOs are absent or underdeveloped.

Read this table like a sliding scale

• Rows 1-2 (voluntary): policymakers can start here to get money flowing quickly without new legislation.

• Rows 3–4 (statutory): move down only if free-riders or repertoire gaps persist.

This graduated menu lets the EU calibrate pressure: support innovation when industry co-operates, but guarantee remuneration where it does not.

Table 12: Graduate menu

Model	Legal form	Typical use today	Pros / Cons for AI-training
1. Voluntary Blanket Licence	Pure contract between AI firm & sector CMO	Pan-EU radio, some image banks	✓ Quickest to implement; ✗ Covers only willing parties, excludes unrepresented sectors
2. Extended Collective Licence (ECL)	National law deems CMO licence to cover non-members (opt-out possible)	Nordic TV catch-up; CDSM Art. 8 (out-of-commerce)	✓ One-stop shop; ✗ Requires opt-out portal, notice infrastructure, and cross-border coordination
3. Flat Levy	Statutory levy, disbursed by public or sectoral fund	Private-copy levies on devices/media	✓ Easy to collect; ✗ Not usage-linked, may face WTO/TRIPS compatibility concerns
4. Statutory Exception + Equitable Remuneration *(recommended baseline)*	EU-level exception + unwaivable right, administered by CMOs	Cable retransmission; resale right; DSM Art. 18	✓ Legally robust, scalable, fair; ✗ Requires EU legislation, high CMO audit capacity

While the graduated toolbox offers policymakers a flexible path, this study ultimately recommends jumping straight to Row 4 – a statutory exception coupled with an unwaivable right of equitable remuneration, administered by CMOs.

• Legal certainty: it eliminates doubt about the legality of training and the enforceability of payments.

• Efficiency: one mandatory scheme is cheaper for developers than negotiating dozens of voluntary blanket licences.

• Fairness & alignment: it mirrors existing EU solutions (cable re-transmission, resale right, DSM Art. 18) and merges with the AI Act's transparency duties.

• Fail-safe: voluntary pilots (Rows 1–2) can still run in parallel for early adopters, but if they leave coverage gaps the statutory back-stop guarantees that all creators receive a share.

E) Data-driven allocation of remuneration

Effective distribution of remuneration hinges on traceability—but full, itemised tracking of training data inputs is unrealistic at scale. Instead, **a data-driven approach grounded in transparency and metadata** offers a workable compromise. Under this model, disclosures mandated by the AI Act (Art. 53) are combined with existing infrastructure—so-called **metadata hubs**, such as ICE (International

Copyright Enterprise)[351] for music or Mint for audiovisual works[352]—which aggregate rights data from multiple jurisdictions. These hubs allow **probabilistic allocation**: identifying likely matches between training data categories and rightsholders repertoires. Where precise matching is not feasible, fallback mechanisms ensure inclusiveness, for example by supporting underrepresented creators or applying proportional distribution keys. A phased rollout would focus initially on metadata-rich sectors like music and visual arts before expanding more broadly.

1) AI developers would be required under the AI Act (Art. 53) to publish detailed summaries of training data sources.

2) Metadata hubs (e.g. ICE, Mint) and national registries would support probabilistic and statistical allocation of revenues.

3) Where matching is not possible, fallback distribution methods would ensure fair and inclusive allocation (e.g. support to emerging creators).

4) Sectoral Rollout: Given differing levels of readiness among creative sectors, the model could initially apply to music and visual works, expanding over time.

The proposed statutory right to equitable remuneration for the use of copyrighted works in AI model training must be supported by a credible and transparent methodology for apportioning value across heterogeneous datasets. AI training data is often compiled from diverse sources, including literary texts, encyclopaedias, academic journals, software code, press content, and user-generated material. These datasets differ widely in their commercial value, creative density, and frequency of reuse. A uniform distribution of remuneration across all contributors would be inefficient and arguably inequitable.

To ensure proportionality and administrative feasibility, a **hybrid allocation model** could be adopted, based on the following principles:

1. Token-based proportionality with content-type weighting

Remuneration could be allocated using a multi-factor formula that includes:

- **Token share:** The number of tokens (words, image pixels, or audio frames) associated with a given dataset, as a share of the total training corpus.
- **Content-type multipliers:** Higher weights for categories such as journalistic content, professional photographs, scientific publications, or curated audio-visual scripts.
- **Usage impact indicators:** Where available, logs or metadata indicating downstream reuse (e.g. through fine-tuning stages, API frequency, or citations in output) could inform impact-based redistribution.

[351] ICE is a global copyright database developed by PRS (UK), STIM (Sweden), and GEMA (Germany). See: https://www.iceservices.com

[352] Mint is the metadata infrastructure used by the International Federation of Film Archives (FIAF). See https://mintproject.github.io/mint/

This type of **probabilistic distribution model** is widely used in other copyright domains where direct tracking is not feasible.[353]

2. CMO-based implementation and oversight

Collective Management Organisations (CMOs) are best placed to administer the remuneration scheme. However, to ensure consistency, transparency, and accountability, the following governance measures are recommended:

- The proposed **AI & Copyright Unit** within the EU AI Office (see Section 4.1) should be empowered to **review and provide oversight on the methodologies or allocation criteria used by CMOs**, particularly in relation to AI training compensation schemes.
- CMOs should comply with **Articles 12 to 16 of the Collective Rights Management Directive (2014/26/EU)** regarding transparent distribution, rightsholders information, and fair deduction of management fees.[354]
- A **model distribution formula** could be developed in cooperation with EUIPO and adopted across Member States.

3. Precedents in EU copyright practice

Similar methodologies have already been implemented successfully in other areas of collective licensing:

- **Cable retransmission (Directive 93/83/EEC):** where remuneration is allocated using proxy indicators such as channel weightings and audience share.[355]
- **Public Lending Right (PLR):** which uses public library lending data as a basis for author compensation.[356]
- **Private copying levies:** where statistical sampling and market studies inform distribution.[357]

[353] See e.g. European Commission, Study on Emerging Issues in Collective Licensing Practices, 2021. Available at https://op.europa.eu/publication-detail/-/publication/8768f709-4c15-11ec-91ac-01aa75ed71a1; see also Jiachen T. Wang et al., *An Economic Solution to Copyright Challenges of Generative AI*, arXiv (Apr. 2024), https://arxiv.org/abs/2404.13964 (illustrating a probabilistic, game-theoretic framework for allocating royalties to copyright holders based on their data's contribution to AI-generated content, using Shapley values to estimate proportional value in cases where direct tracking is infeasible).

[354] Directive 2014/26/EU of the European Parliament and of the Council of 26 February 2014 on collective rights management and multi-territorial licensing of rights in musical works for online use in the internal market.

[355] Council Directive 93/83/EEC of 27 September 1993 on the coordination of certain rules concerning copyright and rights related to copyright applicable to satellite broadcasting and cable retransmission, O.J. (L 248).

[356] Directive 2006/115, of the European Parliament and of the Council of 12 December 2006 on Rental Right and Lending Right and on Certain Rights Related to Copyright in the Field of Intellectual Property, 2006 O.J. (L 376) 28 (EC) ("States shall provide, subject to Article 6, a right to authorise or prohibit the rental and lending of originals and copies of copyright works, and other subject matter as set out in Article 3(1)."). See e.g. Jim Parker, The Public Lending Right and What It Does, World Intell. Prop. Org. [WIPO] Mag. (June 2018). Available at https://www.wipo.int/en/web/wipo-magazine/articles/the-public-lending-right-and-what-it-does-40437.

[357] However, some studies question the efficiency and overall welfare impact of copyright levies, especially in digital contexts: see Martin Kretschmer, Private Copying and Fair Compensation: An Empirical Study of Copyright Levies in Europe – A

These examples demonstrate that **efficient and equitable remuneration is possible even in the absence of granular usage tracking**, provided that distribution frameworks are transparent, periodically audited, and based on agreed proxies. In addition, the European Union has already used private international law mechanisms to prevent the concurrent application of multiple national laws within its member states. This is achieved through directives mandating the adoption of a single governing law, specifically aimed at streamlining copyright clearance and enhancing the freedom to provide services on a multi-territorial basis.[358]

Policy Recommendation

To meet the requirement of appropriate remuneration while avoiding excessive administrative burdens on developers or CMOs, the European Parliament should consider:

1. Mandating the use of **token-weighted, category-adjusted distribution formulas** for AI-related remuneration in consultation with CMOs, rightsholders groups, and academic experts in digital copyright metrics.
2. Establishing a **technical working group** under a AI & Copyright Unit (within the EU AI Office) to define acceptable proxies, validate reporting standards, and provide oversight.
3. Requiring **CMOs to publish annual summary reports** detailing distribution methodologies and usage data.

F) Enforcement and compliance system

Ensuring the practical implementation of the remuneration framework requires a proportional and flexible enforcement architecture. This system must be robust enough to guarantee compliance by major AI actors, while also scalable and accessible for smaller developers. The following multi-level toolkit outlines concrete enforcement levers—ranging from reporting duties and audit rights to institutional oversight and soft compliance mechanisms—that together support legal certainty, fairness, and operational efficiency.

1) Reporting obligations for AI developers to declare categories and sources of training data;
2) Dataset brokers or curators that supply training corpora shall be deemed "providers" when they make protected material available for AI training, and must file the same reports and remuneration declarations;

3) CMOs' authority to collect, audit, and distribute funds based on metadata and probabilistic models;

Report for the UK Intellectual Property Office (2011). Available at https://ssrn.com/abstract=2710611; Christian Peukert, Copyright Levies and Cloud Storage: Ex-Ante Policy Evaluation with a Field Experiment, 53 Research Policy, 1-12 (2024).

[358] See e.g. Council Directive 93/83/EEC, cit. and Directive 2010/13/EU, of the European Parliament and of the Council of 10 March 2010 on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services, O.J. (L 95/1).

4) Oversight by national or EU copyright bodies—acting under existing legal frameworks such as the InfoSoc Directive and the 2004 Enforcement Directive—may include administrative fines or other sanctions for non-compliance;[359]

5) Access to judicial or ADR mechanisms for creators to challenge misallocation or non-payment;

6) Gradual rollout starting with sectors that already have mature metadata systems;

7) Soft enforcement mechanisms in the initial stages, including voluntary certification schemes, nudges, and compliance incentives to encourage uptake without litigation;

8) A central EU-level clearinghouse for small and medium AI developers, modelled after the One-Stop Shop (OSS) system used in VAT reporting, to streamline declarations and payments while reducing administrative burdens.

9) Micro- and small enterprises whose cumulative training compute does not exceed a threshold (e.g. 500 GPU-hours/year) may opt for a simplified lump-sum tariff administered by the OSS clearinghouse. Finally, a layered enforcement framework will be essential to make these remuneration rights operational. Courts should be equipped to grant injunctive relief against models trained on infringing datasets, based on the logic of Article 11 of the Enforcement Directive (2004/48/EC). In parallel, administrative enforcement pathways—including an expanded role for the AI Office or national authorities—could provide a faster, less burdensome route for addressing compliance failures, particularly in relation to transparency and opt-out violations. Lawmakers might also consider statutory damages or presumptive remedies for unauthorized use of protected content, helping to lower the evidentiary burden for individual rightsholders. Without these tools, the proposed licensing and remuneration framework risks becoming de facto unenforceable, especially for smaller creators and independent artists. Effective enforcement is not only a matter of legal coherence—it is also a prerequisite for preserving the credibility of the EU's digital rights framework and ensuring that economic justice is meaningfully achievable in the age of generative AI.

G) Feasibility and political viability

The success of any remuneration model depends not only on legal soundness, but also on practical feasibility and stakeholder alignment. The table below summarises the strengths and limits of the proposed mechanism across key implementation dimensions—legal, administrative, technical, and political. It is followed by a short stakeholder landscape analysis and a comparative advantages summary to support informed policy design.

Table 13: Strengths and limits of the proposed mechanism

Component	Status	Feasibility Rationale

[359] Directive 2004/48/EC of the European Parliament and of the Council of 29 April 2004 on the enforcement of intellectual property rights ("Enforcement Directive"), OJ L 157, 30.4.2004, 45–86 (establishing measures, procedures, and remedies necessary to ensure the enforcement of intellectual property rights, including injunctions, damages, and evidence-gathering provisions).

Legal Basis	✓ High	Existing EU analogies (InfoSoc, DSM, Resale Right); compatible with Berne three-step test and WTO/TRIPS.
Administrative Infrastructure	⚠ Mixed	Strong in music/audiovisual, weaker in text/image — phased rollout recommended
Technical Feasibility	✓ High	Metadata hubs already exist (ICE, Mint); AI Act mandates transparency
Political Viability	✓ Moderate	Balanced between innovation and creator protection; avoids full licensing model; requires careful SME carve-outs
Enforcement & Compliance	⚠ Moderate	Relies on scaled-up regulatory capacity; cross-border consistency depends on EU-level clearinghouse and harmonised audit mechanisms
Implementation Risks	⚠ Present	Requires buy-in from large AI developers; potential regulatory capture or CMO underperformance; mitigated through oversight and inclusive governance

Creators & CMOs: Strong support expected; the scheme mirrors positions taken by music and visual-arts sectors in recent Commission consultations.

Large AI developers: Likely push-back on an unwaivable obligation, yet a statutory exception plus collective remuneration is still cheaper and legally clearer than negotiating bespoke licences for billions of files.

Member States: Countries with robust CMO ecosystems (e.g. France, Germany, Spain) are natural allies; a phased roll-out and SME carve-outs can win over more digital-first or sceptical jurisdictions (e.g. Estonia, Ireland, Sweden).

Comparative Advantages

- Legal Legitimacy: Builds on existing, accepted EU copyright tools.
- Operational Feasibility: Uses infrastructure already in place (CMOs, AI Act transparency, metadata hubs).
- Fairness: Equitable distribution even in absence of perfect traceability.
- Scalability: Sector-sensitive and adaptable to future AI models.
- Policy Alignment: Integrates AI Act principles, DSM Art. 18, and broader EU copyright strategy.

Implementation, Enforcement, and Legal Coherence: Despite its advantages, the proposed model requires careful attention to three overarching risk areas. **First**, implementation challenges—such as resistance from AI developers, unequal sectoral representation in CMOs, and risks of regulatory capture—should be mitigated through transparent oversight and inclusive governance structures. **Second**, cross-border enforcement remains a significant hurdle; consistent application across Member States will depend on effective standardisation, interoperable audit tools, and central coordination—such as an EU-level clearinghouse. **Third**, compliance with international norms is critical. By conditioning the statutory exception on equitable remuneration, the proposal seeks to comply with the Berne Convention's three-step test, particularly the requirement that exceptions must not conflict with the normal exploitation of works. This model mirrors mechanisms—such as private copying levies and

the artist's resale right—that have already passed scrutiny under both EU and WTO/TRIPS frameworks, providing a legally strong path to rebalancing creative value chains in the AI era.

While introducing statutory remuneration schemes is essential to maintain fairness towards creators, policymakers must also carefully consider the compliance costs and practical burdens, particularly for SMEs and start-ups. To mitigate disproportionate impacts, graduated obligations based on company size and revenues, simplified licensing procedures, or threshold-based exemptions should be explored. Additionally, open-source and research-driven initiatives should be supported through tailored regulatory carve-outs, ensuring that fairness in remuneration does not inadvertently stifle European innovation ecosystems.

H) Sample legislative amendment and standardization

In order to operationalise the proposed statutory right to remuneration for the use of copyrighted content in AI training, legislative adjustments should be accompanied by technical standardisation measures that ensure both enforceability and proportionality. The following sample amendment outlines a possible formulation under EU law, complemented by a roadmap for implementing metadata-based opt-out mechanisms in line with Article 4 of the Directive on Copyright in the Digital Single Market (CDSM).

Sample Legislative Amendment (Model Clause)

Article XX – Use of protected content in AI model training

1. Notwithstanding Articles 2 and 3 of Directive 2001/29/EC, the use of lawfully accessible works and other subject matter for the sole purpose of training generative artificial intelligence systems shall be permitted, provided that such use is accompanied by a fair and proportionate remuneration to the relevant rightsholders.

2. For the purposes of this Article, "training" includes initial training, re-training, fine-tuning, or any process in which protected works are ingested to adjust model parameters. It excludes the inference stage in which end-users interact with a pre-trained model.

3. The right to remuneration shall be unwaivable and exercised collectively through collective management organisations designated by the Member States.

4. The amount of remuneration shall take into account the scale of use, the nature of the works used, and the commercial value of the resulting AI system or model.

5. Providers of generative AI systems shall submit reports indicating the general categories, types, and sources of data used, in accordance with Article 53 of Regulation (EU) 2024/1689 (AI Act).

6. Member States shall ensure that appropriate procedures are in place for the distribution of remuneration to rightsholders, including fallback mechanisms in cases of unverifiable use.

7. Where the rightsholder has embedded a machine-readable opt-out signal in accordance with technical standards adopted under this Article, such content shall not fall under the obligation in paragraph 1, unless the metadata has been removed or ignored without justification.

8. The Commission shall be empowered to adopt implementing acts specifying the format, interoperability requirements, and technical means for communicating such opt-outs.

To ensure the enforceability of opt-outs the development of metadata-based opt-out infrastructure should leverage existing technical standards. These include:

- W3C DCAT[360] and RightsML[361] for data cataloguing and rights expression;

- IPTC's Digital Source Type[362] and Rights Expression Language (rNews)[363];

- ETSI standardisation efforts for machine-readable copyright metadata.[364]

A two-phase roadmap is suggested:

Phase 1 – Voluntary implementation: Support uptake of interoperable opt-out signals by major platforms and dataset providers, including pilot registries hosted by EUIPO.

Phase 2 – Mandated thresholds: Introduce mandatory compliance for GPAI providers exceeding compute or revenue thresholds, with enforcement coordinated by the AI Office.
To address potential burdens on SMEs:

The EU should fund open-source opt-out tagging tools and offer technical support via the Digital Europe Programme;

Phased compliance timelines or sandbox exemptions should be granted to micro-entities;
Public guidelines and templates should be co-developed with standards bodies and creative sector representatives.

I) Comparative analysis and rationale

[360] DCAT (Data Catalog Vocabulary) is a W3C standard for describing and sharing datasets across platforms through interoperable metadata. It supports discoverability and integration of public and private data catalogs. See World Wide Web Consortium (W3C), Data Catalog Vocabulary (DCAT) Version 2, W3C Recommendation, 2020. Available at https://www.w3.org/TR/vocab-dcat-2/

[361] RightsML**,** developed by the IPTC and maintained under ETSI, is a machine-readable rights expression language designed to represent copyright permissions, restrictions, and obligations in a structured XML format. See IPTC & ETSI, RightsML Specification, 2012. Available at https://iptc.org/standards/rightsml/

[362] The Digital Source Type vocabulary, developed by IPTC (International Press Telecommunications Council), is used to classify the origin of digital content (e.g., "user-generated," "professional," or "aggregated"). It enables more precise metadata tagging for content rights and provenance. See IPTC, Digital Source Type Vocabulary, IPTC Documentation. Available at https://iptc.org/news/new-digital-source-type-term-added-to-support-inpainting-outpainting-in-generative-ai/

[363] rNews is a metadata standard also developed by IPTC that applies the schema.org vocabulary to news content, enabling structured, machine-readable information about authorship, licensing, and usage terms. See IPTC, rNews Metadata for News Industry, 2011–2013. Available at http://dev.iptc.org/rNews-1-Introduction-to-rNews

[364] ETSI (European Telecommunications Standards Institute) has developed standards for machine-readable rights expression languages, including RightsML**,** in collaboration with IPTC. RightsML allows the encoding of permissions, prohibitions, and obligations associated with digital content in a structured XML format. These standards aim to support automated copyright compliance in digital and AI ecosystems by enabling interoperability between rights holders, content platforms, and AI systems. See

Table 14: Comparative Overview of Three Remuneration Models for AI training

Model	Legal Basis	Main Mechanism	Distribution Method	Strengths	Weaknesses
1. Statutory Exception + Equitable Remuneration (this proposal)	New copyright exception + unwaivable remuneration right (based on InfoSoc Art. 31, Resale Right, DSM Art. 18)	Statutory training use + collective remuneration via CMOs	Probabilistic + metadata-informed distribution	Legally grounded; scalable; sector-sensitive; aligned with AI Act	Requires legislative change; relies on CMOs' efficiency
2. Levy-Based Remuneration Scheme (Senftleben)	New levy system linked to training activities	Flat-rate levy on AI developers or model usage	Redistribution via public/cultural funds	Conceptually simple; no need for traceability	Weak link to actual usage; limited creator targeting
3. Statutory License via Fundamental Rights Balancing (Geiger & Iaia)	Fundamental rights (freedom of expression, right to culture) justify licensing without consent	Non-voluntary license + possible collective remuneration	Abstract or undefined	Strong rights-based justification; innovation-friendly	Implementation pathway unclear; lacks infrastructure linkage

Why not levies alone? A flat levy on devices or compute bills disconnects payment from actual training intensity and repertoire value; it also risks WTO/TRIPS scrutiny for disguised turnover taxes. *Why not individual licences?* Scale makes them unworkable. The proposed statutory exception + remuneration right **preserves systemic proportionality, complies with EU treaty obligations, and leverages existing CMO infrastructure**—hence it is the most immediately actionable compromise.

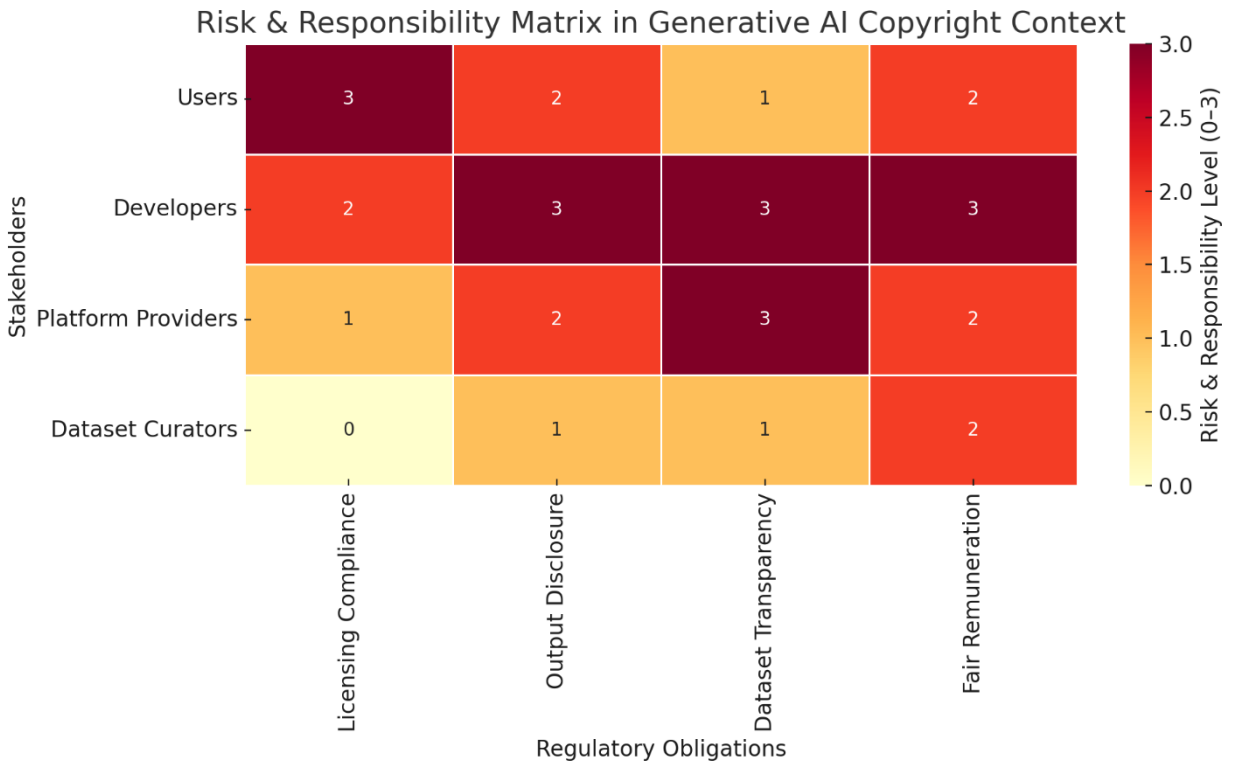
This proposal offers a practical middle ground between copyright enforcement and technological innovation. Unlike flat levies, which disconnect payment from actual usage, or abstract licensing models that lack operational clarity, this system builds on proven EU legal tools. It channels remuneration through CMOs using existing metadata systems, while avoiding unrealistic requirements like granular tracking. It also accounts for the market realities faced by SMEs through a simplified central clearinghouse. It aligns with the EU's risk-based approach to AI regulation, ensuring scalability, fairness, and transparency without imposing unworkable burdens. For policymakers seeking a forward-looking, sector-sensitive, and legally robust solution, this model represents the most actionable path forward.

J) Risk-based allocation of obligations

In order to support the implementation of the proposed remuneration model—and to ensure that it is enforceable, proportionate, and targeted—it is essential to integrate a complementary risk-based framework. This approach aligns with the logic of the EU Artificial Intelligence Act and other digital regulatory instruments, which differentiate duties based on actors' control over risk and their role in the technological ecosystem.

In this context, the equitable remuneration right outlined above cannot exist in a vacuum: it must be embedded in a system that distributes regulatory burdens according to each stakeholder's practical ability to comply and influence outcomes. Developers and platform providers, for example, shape the architecture and deployment of generative AI systems and therefore have greater capacity to implement licensing compliance and dataset transparency than end users. Dataset curators, meanwhile, control the quality and legitimacy of the inputs but are rarely addressed in current debates. The matrix below offers a conceptual tool to visualise how the key copyright-related obligations underlying the remuneration scheme—such as licensing compliance, output disclosure, dataset transparency, and fair remuneration—can be mapped against different stakeholder groups. It reflects a scalable model in which regulatory duties are matched to institutional capability, paving the way for more enforceable and just implementation of the proposed framework.

Figure 2: Risk and Responsibility Matrix in the GenAI Copyright Context



Such a risk-tiered model may assist policymakers in developing targeted transparency requirements, clearer due diligence standards, and equitable remuneration mechanisms. Importantly, it also offers a pathway to distribute the costs of compliance and enforcement more fairly—ensuring that those who derive the greatest economic value from AI systems also bear a corresponding share of the regulatory obligations.

K) Specific Considerations for Open-Source GPAI Models

Open-source General-Purpose AI (GPAI) models represent a unique category within the broader AI ecosystem. Unlike proprietary systems developed and commercialised by large technology firms, open-source GPAIs are typically released under free or permissive licences, allowing public access to the model's architecture, source code, and in some cases, its training datasets and model weights.

These models are widely used by universities, researchers, civil society organisations, and start-ups. Their openness promotes transparency, reproducibility of research, and decentralised innovation. However, their legal status within the evolving EU regulatory framework for AI and copyright remains under-defined. While the AI Act offers conditional exemptions for GPAI models released under free and open-source licences (see Article 2(12)[365] and Article 53(2)),[366] questions persist regarding how such exemptions interact with copyright-based obligations—particularly in relation to training data and authors' remuneration.

Given their growing societal and economic relevance, a nuanced regulatory approach is required—one that preserves the collaborative nature of open-source innovation, while ensuring consistency with fundamental copyright principles and policy goals.

Open-source GPAI models are often distributed under standard software licences, such as **GNU General Public License (GPL) v3**, **Apache License 2.0**, or newer community-drafted licences like **OpenRAIL.**[367] Each presents different implications for compliance with proposed EU-level copyright measures, including the transparency of training data and the introduction of a statutory remuneration mechanism.

Table 15: Differences between 2 standard software licences

Legal Element	**GPL v3 (Copyleft)**	**Apache 2.0 (Permissive)**
Scope of Coverage	Requires that any modified or derivative work be distributed under the same licence terms. Uncertainty exists as to whether model weights or fine-tuned variants fall under this obligation.[368]	More flexible: only modified code files require preservation of attribution; model weights can be redistributed under separate terms.
Patent and IP Clauses	No explicit patent licence is granted. This may create legal uncertainty for downstream users.	Includes an express patent grant, reducing IP-related legal risks for developers and users.
Licence Compatibility	Incompatible with some community-specific or field-of-use restricted licences. Stacking of obligations may inhibit reuse.	Typically compatible with layered licensing frameworks and easier to reconcile with copyright-compliant obligations.

[365] Article 2(12) AI Act provides that the Regulation "does not apply to AI systems released under free and open-source licences, unless they are placed on the market or put into service as high-risk AI systems or as an AI system that falls under Article 5 or 50."

[366] Article 52(2) Ai ACT providing thet "the obligations set out in paragraph 1, points (a) and (b), shall not apply to providers of AI models that are released under a free and open-source licence that allows for the access, usage, modification, and distribution of the model, and whose parameters, including the weights, the information on the model architecture, and the information on model usage, are made publicly available. This exception shall not apply to general-purpose AI models with systemic risks."

[367] See Danish Contractor and Carlos Muñoz Ferrandis, BigScience Large Open-science Open-access Multilingual Language Model (2022). Available at https://bigscience.huggingface.co/blog/the-bigscience-rail-license

[368] See e.g. Pamela S. Chestek, A Promise Without A Remedy: The Supposed Incompatibility of The GPLV2 and Apache V2 Licenses, 40 Santa Clara High Tech. L.J. 303 (2024).

Legal Element	**GPL v3 (Copyleft)**	**Apache 2.0 (Permissive)**
Reuse in Commercial Environments	Complex, as copyleft provisions may discourage commercial entities from integrating GPL-licensed models into proprietary products.	Often favoured in commercial contexts due to minimal licensing restrictions.

Policy Implication: The compatibility of open-source licences with EU copyright and AI regulation varies. Permissive licences, such as Apache 2.0, are more naturally aligned with the proposed model of collective licensing and statutory remuneration. By contrast, strong copyleft licences like GPL v3 may raise ambiguity around the legal status of derivative works (e.g., model weights) and their eligibility for inclusion under a new regulatory mechanism.

The AI Act's Article 53 introduces a requirement for GPAI providers to publish "detailed summaries" of training data. For centrally developed and commercially maintained systems, compliance with this obligation is relatively straightforward. For open-source models, however, the decentralised and collaborative nature of development introduces challenges:

- Lack of a clearly identified "provider" or controller: Many community-developed GPAI models are maintained by ad hoc collectives or research networks. There is no single legal entity to bear the compliance burden.

- Partial or incomplete dataset documentation: Large-scale open-source models often rely on hundreds of public or semi-public datasets. These datasets may lack consistent metadata, licensing information, or copyright provenance, making summary creation resource-intensive and legally risky.

- Evolving architecture and frequent forking: Open-source models evolve rapidly through community "forks" and contributions.[369] Tracking training inputs and dataset modifications across forks requires sophisticated version control mechanisms and substantial coordination, which may exceed the capacity of non-profit developers.

Feasibility Recommendation:

In order to address these structural limitations while preserving transparency:

- The "detailed summary" requirement should be modular and scalable. A tiered disclosure model could be adopted:

- Core datasets accounting for the majority of training volume should be identified individually.

- Secondary datasets may be aggregated or listed in summary format.

- Clear disclaimers should be permitted where provenance cannot be verified.

[369] Open-source models frequently evolve through community contributions and forks, leading to rapid architectural changes. This iterative process is evident in modern ecosystems. See Linus Nyman & Juho Lindman, Code Forking, Governance, and Sustainability in Open Source Software, 3 Tech. Innovation Mgmt. Rev. 7, 9–12 (2013).

• An EU-backed template for dataset summaries, co-developed with the AI Office and relevant standards bodies (e.g., W3C, IPTC), should be adopted to facilitate consistent and cost-effective compliance.

• For collaborative open-source projects, the role of "provider" under the AI Act should be attributed to the lead maintainer or initial releasing entity, as is common in software governance.

To avoid unintentionally burdening small research groups or non-commercial developers, a form of proportional relief could be built into both the transparency and copyright-remuneration frameworks.

We recommend the following tiered approach:

• Threshold-based exemptions: Forks that (i) are developed by non-profit entities, and (ii) remain below a defined compute or revenue threshold (e.g. <250 GPU-hours or <€750,000 turnover) should benefit from simplified obligations:

- Short-form summaries of training data.

- Fixed low-cost contributions to collecting societies rather than per-use remuneration.

- Exemption from metadata fingerprinting or documentation duties.

• Graduated escalation: Once a fork is integrated into a commercial service, or deployed in a high-risk AI system, full obligations (including standardised dataset summaries and remuneration payments) would apply automatically.

• Optional EU labelling scheme: A "**yellow label**" could signal open, non-profit GPAI projects that operate under simplified compliance. This could encourage responsible innovation while maintaining legal certainty for downstream users.

Open-source General-Purpose AI (GPAI) models have become a foundational component of the European AI research and innovation ecosystem. Their specific licensing structures and decentralised development practices demand tailored regulatory approaches.

To ensure that EU copyright law and AI regulation foster innovation without imposing disproportionate compliance burdens, the following elements are critical:

1) Legal clarity on licence interaction: The statutory remuneration scheme should be explicitly compatible with permissive open-source licences, while also addressing potential conflicts with strong copyleft models.

2) Scalable compliance mechanisms: Article 53 transparency obligations must be attainable for decentralised or volunteer-led initiatives. Standardised templates, tiered disclosure obligations, and public repositories should be developed to enable compliance at scale.

3) Proportionality and inclusivity: Relief mechanisms for small-scale or non-commercial forks are essential to sustaining Europe's leadership in open and ethical AI development.

4) Governance and enforcement: The proposed AI & Copyright Unit within the EU AI Office (see Section 4.1) should be tasked with assessing also the compliance status of open-source projects and offering

guidance or informal mediation in disputes between copyright holders and model developers—particularly in grey areas involving derivative works or dataset licensing.

Table 16: Three-Pillar Check

Pillar	Status	Why?
Epistemic	🟡	Relies on AI Act disclosures and CMO metadata hubs; however, direct audit rights for individual creators remain undefined.
Normative	🟢	Establishes a statutory exception paired with an unwaivable right to equitable remuneration—ensuring revenue-sharing with rightsholders.
Systemic	🟡	Envisions CMO oversight and penalties, but the proposed EU-level clearinghouse is not yet operational or institutionally anchored.

Note: Turns the Normative light green; Epistemic & Systemic become green once 4.3's audit tools are in place (see Sections E, F, and I).

4.4. Clarify protection status of AI-assisted vs AI-created works

A) Exclude AI-only outputs from copyright protection

The question of whether and under what conditions content generated with the assistance of artificial intelligence qualifies for copyright protection lies at the heart of current legal uncertainty surrounding generative AI. While EU copyright law maintains a clear human-centric approach to authorship—requiring "the author's own intellectual creation" as defined by the CJEU—the emergence of generative models capable of algorithmically assembling expressive elements highlights the need for clearer boundaries between non-protectable machine-derived outputs and protectable human–machine co-authored works.

It is essential to reaffirm that purely AI-generated outputs—those produced without any human creative input—do not meet the originality threshold required for copyright protection under EU law. This stance is already consistent with CJEU jurisprudence and reflects the foundational principle that protection is reserved for human intellectual creation.

The entire edifice of copyright law—built on principles such as the distinction between ideas and their expression, the requirement of originality, and the legal notion of authorship—presupposes the involvement of a human creator. In the absence of human intellectual input, there can be no original expression of ideas, and thus no work eligible for protection under copyright law.[370]

Nonetheless, ambiguity persists in practice, as national authorities and creators confront borderline cases involving partial human curation or minimal intervention. To prevent divergent interpretations across Member States and pre-empt legal fragmentation, it is recommended that the European Commission—possibly in cooperation with EUIPO—issue **guidance clarifying that AI-only outputs fall**

[370] See Matt Blaszczyk, Impossibility of Emergent Works' Protection in U.S. and EU Copyright Law, 25 North Carolina Journal of Law & Technology, cit. at 161

outside the scope of copyright protection, and that only works exhibiting significant human creativity may qualify for protection. Such clarification could be included soft law instruments (e.g. Commission Communications, expert group recommendations).

B) Clarify public domain status and regulatory boundaries

This guidance should also address a persistent public misconception: that AI-generated outputs are either automatically protected or entirely unregulated "free goods." In reality, **non-protectable outputs revert to the public domain**, yet may still be subject to other legal regimes (e.g. trade secrets, database rights, personal-data or personality-rights rules, and contractual terms of service). Failing to communicate this nuance risks two opposite—and equally harmful—outcomes: (i) commercial actors may try to over-claim proprietary control over machine outputs, stifling legitimate reuse; (ii) users may unknowingly infringe other rights or regulatory constraints.
To enhance legal certainty while supporting responsible innovation, EU Intitutions—working with EUIPO, the European AI Office, and national IP offices—should adopt a three-pillar communication and labelling strategy:

1. **Authoritative guidance & public info-sheets**

Issue a concise guidance document and multilingual info-sheets that:

- confirm that outputs lacking the requisite level of human creativity enter the public domain by default;

- map out residual regimes that can still constrain reuse (trade-secret law, sui generis database right, consumer-protection or data-protection rules, contractual licences);

- provide real-world examples (e.g. "AI-generated weather data vs. AI-generated brand mascots").

2. **Voluntary EU "AI-Output Labelling Toolkit"**

- Develop a set of machine-readable metadata tags (e.g. ai-output:public-domain, ai-output:restricted, ai-output:personal-data-sensitive) for creators, platforms and model providers.

- Encourage large content hosts, open-source repositories, and GPT-style model gateways to display these badges prominently, improving downstream clarity for SMEs, educators and the cultural sector.

3. **Outreach & help-desk support**

- Launch a targeted outreach campaign (webinars, social-media explainers, sectoral roadshows) aimed at creators, developers and SMEs.

- Establish a help-desk—possibly within the EU IP Helpdesk network—offering first-line advice on reuse of AI outputs, opt-out signals, and conflict checks with other rights frameworks.

This integrated approach will prevent over-enclosure of the digital commons, reduce inadvertent infringement, and foster a culture of transparent licensing and responsible reuse across the EU creative and tech ecosystems.

In addition to legislative responses within copyright law, horizontal coordination with competition and consumer protection authorities may be warranted. As generative content proliferates, the ability of large AI developers to saturate cultural and informational markets with unlicensed, public-domain

outputs could have structural effects on content diversity, pricing, and creator viability. Exploring the interface between copyright exhaustion, public domain status, and market power could thus form part of a broader regulatory strategy—potentially involving DG COMP, consumer law instruments, or sector-specific codes of conduct.

C) Provide criteria for assessing human authorship in AI-assisted creation

In the grey area of AI-assisted creation—where humans interact with AI tools to varying degrees—the need for legal certainty is particularly acute. Current EU law offers no clear test for determining when human involvement crosses the threshold from technical facilitation to genuine authorship. It is therefore recommended that the EU initiate the development of concrete, non-binding criteria or case examples to assess authorship in AI-assisted works. These could include factors such as:

- the selection and refinement of AI prompts with a specific creative intent;

- the human curation and adaptation of AI-generated variants;

- the combination of AI outputs with original human content in a meaningful and non-trivial way.

Such criteria should not adopt a formalistic or numerical threshold but should instead focus on the qualitative aspects of human creative choices and expressive control. This would allow creators, users, and enforcement bodies to navigate the legal landscape with greater confidence while preserving the EU's foundational commitment to human creativity. While case law from the CJEU may ultimately refine these standards, **interim policy guidance from the EU Copyright Contact Committee or a dedicated expert group** could offer much-needed clarity.

D) Address strategic misattribution of AI-generated content

EU policymakers should address the growing practice of strategically presenting fully AI-generated outputs as original human-authored works, particularly in commercial settings. While such outputs fall into the public domain under current EU law, false claims of authorship may distort copyright expectations, undermine legitimate reuse, and contribute to unfair competition. To mitigate this risk, soft-law instruments or sectoral codes of conduct could encourage the disclosure of AI involvement and prohibit misleading attribution, especially in professional and commercial contexts. These measures would enhance transparency, protect the integrity of the public domain, and support fair market conditions.

E) Align AI Act transparency obligations with copyright goals

The study recommends aligning the implementation of the AI Act—particularly Articles 50(4) and 50(5)—with broader copyright transparency objectives. These provisions require deployers of AI systems to disclose when content has been artificially generated or manipulated, particularly in the case of deep fakes and AI-generated text intended to inform the public on matters of public interest. While not part of the copyright acquis, this obligation serves an important ethical and reputational function. **Ensuring consistent and effective implementation across Member States would reduce consumer confusion**, prevent the misattribution of machine-generated works to human creators, and help safeguard the integrity of creative industries. To support this, the European Commission should consider issuing implementation guidance or, where appropriate, an **implementing act pursuant to**

Article 50(7) in conjunction with Article 98(2) of the AI Act, clarifying what constitutes sufficient disclosure across different content types, including visual, textual, and audio formats.

F) Monitor international divergence in AI authorship standards

The EU should closely monitor developments in third countries. While no major jurisdiction currently grants full copyright protection to AI-generated works, diverging trends are emerging (e.g. limited recognition in China via Court decisions, or expansive authorship definitions proposed in some common-law countries). These asymmetries may give rise to **cross-border recognition issues**, particularly in relation to enforcement, licensing, and market access. The EU should consider establishing an **observatory or working group to track international legal developments** and assess the need for reciprocal treatment or clarifying rules regarding the recognition (or non-recognition) of foreign AI-generated rights under EU law.

G) Resist sui generis rights for machine-generated content

The EU should resist the introduction of new exclusive rights for outputs computed by AI systems without meaningful human input. This view is also supported by the European Parliament's 2020 resolution, which recommended that works automatically synthesised by artificial agents should not be eligible for copyright and that any rights should be conferred only to natural or legal persons under well-defined conditions.[371] While some stakeholders argue for limited or "thin" copyright-like rights to incentivise innovation or manage attribution, such protection would be both conceptually unsound and practically harmful.

There are four main reasons for this:

1) It would distort the creative economy and create unfair competition.

Granting IP-like rights to machine-generated outputs risks introducing large volumes of AI-generated content that may alter competitive dynamics and affect the visibility and value of human-authored works. This would devalue genuine human authorship and undermine the economic viability of professions in the cultural and creative sectors. Creators subject to labour, time, and legal constraints would face systemic disadvantage against automated systems that can generate endless volumes of content with minimal cost.

2) It lacks a normative foundation in copyright law.

Copyright is grounded in human intellectual effort, personal expression, and creativity. Machine-generated works lack intentionality, moral perspective, and expressive autonomy. Introducing a new right for non-human creations would break with this fundamental rationale and force courts and regulators to construct artificial distinctions between machine authorship levels—resulting in legal uncertainty and enforcement challenges.

3) It would encourage perverse incentives and enclosure of the public domain.

[371] See European Parliament resolution of 20 October 2020 on intellectual property rights for the development of artificial intelligence technologies, 2020/2015(INI), 2021 O.J. (C 404) 129, at §15.

Recognising even minimal exclusive rights in machine-generated outputs would encourage the large-scale production of synthetic content for the sole purpose of obtaining control over its distribution, access, or reuse. This could lead to the enclosure of digital commons, reduction of freely usable cultural material, and imbalances in platform power or developers with the means to flood content markets.

4) International fragmentation and trade complications.

Recognising sui generis rights in machine-generated content could lead to fragmentation in global copyright enforcement, complicating cross-border licensing, exceptions, and recognition of human authorship. This would undermine legal certainty and create further friction for EU creators operating in international markets.

Instead of creating new rights, the EU should focus on transparency, authorship attribution, and dataset accountability, while reaffirming that only works meeting the originality standard grounded in human creativity are eligible for protection. This will support legal clarity, market fairness, and the long-term legitimacy of the IP system.

H) Preventing regulatory fragmentation across Member States

In order to ensure legal certainty and internal market cohesion, the EU should take proactive steps to prevent divergent national approaches to AI-generated content and authorship criteria. The EU could issue interpretative guidelines—similar to the 2021 copyright guidance on the CDSM Directive—to clarify under which conditions human involvement in AI-assisted creation satisfies authorship requirements. These could be published as a Communication or via the EUIPO Observatory.

Furthermore, the EU could adopt minimum harmonisation provisions through a targeted amendment to the CDSM Directive or in a future AI-and-copyright legislative package. These provisions could define a common baseline for recognising human authorship in AI-supported works, thereby reducing the risk of inconsistent judicial interpretations. In order to support transparency and legal predictability, a regularly updated EU-wide repository of national case law and implementation practices—maintained by the EUIPO or the AI Office—could provide courts, creators, and platforms with a comparative legal reference tool.

I) Regulating the Market Impact of AI-Generated Outputs: Legal & Governance Toolkit

In response to the challenges outlined in Section 3.3—particularly the **substitution effects and disruption of creative value chains resulting from the widespread use of generative AI**—a complementary set of output-side measures is necessary. To reinforce systemic accountability, the following measures are proposed:

1. Enhance Transparency and Traceability of AI Outputs

Legal vehicle: delegated act under Article 53(6) of the AI Act.

Obligation: within 12 months of the act's entry into force, providers of general-purpose generative models must embed a tamper-resistant, C2PA-compatible watermark or metadata string in every

public-facing output.[372] The schema must (i) identify the model/provider, (ii) state whether the content was human-prompted or fully autonomous, and (iii) remain machine-readable after standard platform compression or format conversion. Failure to comply constitutes a misleading commercial practice subject to Directive (UE) 2019/2161 fines (up to 4 % global turnover).

2. Pilot Output-Linked Remuneration Schemes

Legal vehicle: extended collective-licensing (ECL) pilots authorised by Article 12 of the Collective Rights Management Directive.

Scope: two-year pilots in the music and stock-image sectors, where reliable usage metrics already exist. Commercial GPAI providers whose outputs reach a defined "substitution threshold" (e.g. ≥5 % market share in a content category) pay a levy calculated as a small percentage of EU-derived AI-content revenue. Levies are collected and distributed by CMOs on a token-weighted, content-type-adjusted basis. A sunset-review clause assesses economic impact and decides on EU-wide roll-out or sectoral expansion.

3. Activate the AI & Copyright Unit for Continuous Governance

As mentioned in Section 4.1, the creation of an **AI & Copyright Unit** within the EU AI Office should serve not only as an audit and compliance mechanism, but also as a forum for structured, cross-sector collaboration. This includes regular consultations with creators, AI developers, CMOs, and civil society to monitor the evolving market and legal impact of generative AI, support consistent enforcement across Member States, and promote convergence on technical and ethical standards relevant to copyright and related rights.

Deliverables:

-Quarterly multi-stakeholder fora bringing together creators, GPAI providers, CMOs, consumer groups and competition authorities.

-Annual "Substitution Index" dashboard measuring the traffic share of AI-generated versus human content across major EU platforms; the index informs levy-rate adjustments and triggers competition-policy alerts where necessary.

-Guidance notes on watermark robustness thresholds and best-practice templates for licence reporting.

Table 17: Three-Pillar Check

Pillar	Status	Why?
Epistemic	🔴	No provenance or dataset transparency measures proposed in relation to authorship classification.
Normative	🟢	Clearly affirms public domain status for AI-only outputs and provides criteria for human authorship.

[372] https://c2pa.org/

Pillar	Status	Why?
Systemic	🟡	Recommends EU-level guidance, but lacks binding instruments or enforcement mechanisms across Member States.

Note: Epistemic accountability could be significantly enhanced if paired with transparency tools proposed in Section 4.4, such as watermarking and dataset observatories.

4.5. Support safeguards and content traceability

A) Promoting technical safeguards for content protection

As the deployment of generative AI systems accelerates across sectors, the need for robust safeguards and traceability mechanisms becomes central to any future-proof regulatory approach. While legislative instruments such as the AI Act lay the foundation for transparency and accountability in high-risk AI systems, they do not yet offer a complete framework for ensuring that AI-generated content respects existing copyright obligations or can be effectively traced and monitored. To complement this regulatory baseline, the EU should actively promote the development and uptake of technological tools, collaborative governance models, and legal clarifications that enable both innovation and the protection of creative rights. However, even the most advanced traceability mechanisms—such as watermarking and fingerprinting—cannot, on their own, address the deeper structural risk of economic displacement. Unless targeted interventions are made at the level of **distribution access and algorithmic promotion**, there is a risk that human-created works, although traceable, will remain **invisible or commercially sidelined**. To correct this imbalance, the EU should explore **quota-based content prioritisation** or **visibility guarantees** for human-authored works—drawing inspiration from established instruments in audiovisual media law that safeguard cultural diversity and democratic pluralism.

In parallel, it remains important to support the development and deployment of **technical safeguards** in generative AI systems that reduce the risk of copyright infringement. This includes encouraging innovation in watermarking, fingerprinting, and output filtering techniques. Watermarking can involve embedding invisible metadata or cryptographic markers into AI-generated content, enabling traceability back to the producing model or, where feasible, to the type of source data used. Similarly, improved algorithmic design—such as techniques to prevent verbatim memorization of training data—can reduce the likelihood of infringing outputs being generated. The European Institutions should consider **funding collaborative research and standardisation initiatives** involving academia, industry, and civil society to develop shared benchmarks and interoperable tools that enhance AI model safety in the context of content creation.

Additionally, to ensure equitable access to such tools, the EU should encourage the development of open-source watermarking and fingerprinting technologies. This would allow smaller developers and public institutions to implement safeguards without facing prohibitive licensing or vendor lock-in barriers.

B) Enhancing dataset transparency via opt-out signals

While this study has argued that the opt-out mechanism under Article 4 of the CDSM Directive is not a sufficient legal safeguard for large-scale generative AI training, its technical implementation still offers opportunities to enhance content traceability in the short term. In parallel with the establishment of an EU-wide registry of reservations, embedding opt-out flags directly into commonly used content platforms could reduce the burden on individual creators and ensure broader uptake. These machine-readable signals, based on interoperable protocols, could be automatically indexed by AI developers and integrated into training dataset management systems. Such a solution would not resolve the deeper legal misalignments discussed in Section 4.1, but could serve as a stopgap measure to mitigate unauthorized ingestion and foster a culture of responsible dataset sourcing.

Although there is widespread recognition that standardisation around Article 4(3) opt-outs has failed, few legal scholars have undertaken a systematic review of the technical protocols currently available. Some mention possible technologies, but often without detailing how they actually work. On the other hand, technical experts frequently propose solutions that do not meet the legal requirements for a valid reservation.[373] This disconnect raises a key question: which technologies can truly be considered "machine-readable" within the meaning of Article 4(3) CDSMD? And more importantly, which technologies are durable and future-proof enough to ensure meaningful compliance? Answering this requires an interdisciplinary approach—one that combines technical understanding with legal analysis.

Platforms particularly relevant for implementation include stock image libraries, social media platforms, academic publishing systems, and content management services (CMS). Early cooperation with these actors could significantly improve the visibility and adoption of opt-out signals among creators. Accordingly, without prejudging broader reform, the EU should mandate the creation of a harmonised, machine-readable standard for opt-out signals under Article 4(3) CDSMD—co-developed by legal and technical experts—and make its adoption a condition for lawful dataset collection by AI developers.

C) Strengthening global enforceability of transparency requirements

In order to enhance the effectiveness of transparency requirements, particularly in a global AI landscape dominated by non-EU actors, the European Union should pursue a multipronged strategy grounded in enforceable mechanisms and international cooperation:

First, the EU should actively promote bilateral and multilateral agreements with key AI-developing countries (e.g., the U.S., Japan, Canada, Korea) that establish shared minimum standards for dataset disclosure, model traceability, and auditability. These agreements could draw on existing frameworks, such as the Trade and Technology Council (TTC) and the OECD AI Principles,[374] and should include enforceable provisions on transparency and access to information, with specific reference to copyrighted content used in training. As discussed in Section 2.5, the effectiveness of such

[373] See Hanjo Hamann, Artificial Intelligence and the Law of Machine-ReadabilityA Review of Human-to-Machine Communication Protocols and their (In)Compatibility with Article 4(3) of the Copyright DSM Directive, 15 JIPITEC 102-121 (2024) (systematically reviewing the main human-to-machine communication protocols (robots.txt, meta tags, HTTP headers, etc.) and assessing their (in)compatibility with the legal requirements of Article 4(3) CDSMD).

[374] See US-EU Trade and Technology Council (TTC), https://digital-strategy.ec.europa.eu/en/factpages/eu-us-trade-and-technology-council-2021-2024; OECD AI Principles. Available at https://oecd.ai/en/ai-principles

transparency requirements ultimately depends on their integration with enforceable rights under EU copyright law—particularly the existing, non-waivable obligations established in Articles 18 and 19 of the CDSM Directive.

Second, in the absence of binding global frameworks, the EU should make use of its internal market leverage to impose conditions on the import and deployment of high-risk or general-purpose AI systems. For example, access to the EU market could be conditioned upon the submission of standardised dataset documentation—compliant with Article 53(1)(c) of the AI Act—and subject to randomised third-party audit rights or ex post verifiability assessments. Such a model would mirror the EU's established approach in other digital regulations (e.g., the GDPR's adequacy framework or the DSA's obligations for very large platforms), and could be operationalised through delegated acts adopted under the AI Act.

Third, the EU could mandate that commercial deployers of high-risk or general-purpose AI systems—regardless of where the model is developed—contractually require upstream developers to disclose training data summaries and provenance information. This would ensure a chain of accountability even when the model was trained outside EU jurisdiction. In parallel, EU-funded AI research and public procurement contracts should include transparency-by-design clauses, helping set a de facto industry standard for responsible training practices.

Finally, to foster trust and practical enforceability, the EU could support the creation of an independent, international AI dataset observatory or registry— ideally anchored within existing multilateral institutions such as the OECD or UNESCO, which already host digital policy cooperation frameworks, thereby increasing the feasibility and legitimacy of such an observatory—tasked with collecting and curating disclosures, best practices, and audit methodologies related to generative AI training. Such a body could act as a reference point for regulators, researchers, and rights holders globally, while facilitating convergence around dataset transparency in the creative economy. To incentivise participation from non-EU actors, the EU could link registry cooperation to benefits such as eligibility for research funding, fast-track certification under the AI Act, or access to harmonised assessment tools.

D) Preserving fundamental rights in filtering systems

These safeguards should be developed with a clear commitment to preserving **freedom of expression and lawful uses**, particularly those protected under copyright exceptions such as quotation, parody, and pastiche. To that end, any filtering or moderation mechanism must be calibrated to avoid excessive over-blocking. One useful precedent lies in the content moderation infrastructure developed to combat the dissemination of illegal material such as child sexual abuse images: hashed databases can be employed to identify known works without scanning or restricting lawful expression. A similar logic could be adapted for copyrighted works, balancing enforcement with proportionality. A useful cautionary precedent lies in YouTube's Content ID system, which, despite being one of the most well-known large-scale content recognition infrastructures, has been widely criticized for over-blocking and

discouraging lawful uses such as criticism, commentary, and parody.[375] This experience highlights the risk of overly aggressive filtering and underscores the importance of proportionality, transparency, and effective appeal mechanisms in the design of AI-assisted enforcement tools. Future systems should be built with a strong commitment to safeguarding lawful expression and access to knowledge.

E) Clarifying liability in AI-generated content

Greater **clarity on liability rules** is needed as AI tools increasingly enable user-generated content that may infringe copyright. While platform liability is already addressed under the Digital Services Act (DSA) and, in specific cases, Article 17 of the DSM Directive, the application of liability principles to AI systems—particularly general-purpose models used for creative purposes—remains unclear. To avoid placing undue burden on AI developers, the regulatory framework should distinguish between **tool misuse by users** (where the primary liability lies with the user) and **systematic negligence or facilitation by developers** (e.g. failure to implement safeguards or comply with transparency obligations). The framework should also differentiate between proprietary AI systems with controlled deployment pipelines and open-source or decentralized models, where liability may need to follow different accountability chains. Tailored provisions may be necessary to prevent overregulation of non-commercial or community-based AI projects.

Although the AI Act has now been adopted, the European Union should consider issuing **supplementary interpretative guidance or accompanying soft-law instruments** clarifying that content safeguards developed under Article 50 and related provisions must be implemented in a manner consistent with EU copyright law, including its exceptions and limitations. Such clarification would help ensure that filtering mechanisms do not inadvertently suppress lawful, exception-based, or educational uses, thus preserving the EU's broader commitment to access to knowledge and freedom of expression. In parallel, the Commission should monitor emerging enforcement practices and be prepared to propose targeted updates to the copyright acquis or relevant digital legislation if systemic inconsistencies arise.

F) Integrating AI detection into existing platforms

The EU should promote the **integration of AI-output detection mechanisms into existing content recognition systems**, particularly on platforms that host or disseminate user-generated content. Building on the infrastructure developed under Article 17 of the DSM Directive, AI-generated outputs—particularly those closely mimicking protected works—should be detectable through digital fingerprints, metadata, or identifiable stylistic patterns. These tools can support rights holders in identifying unauthorised uses of their work, and enable platforms to take action in accordance with notice-and-action procedures, while also allowing for human review to avoid erroneous blocking. To support interoperability, the EU should encourage the creation of common technical standards or certification schemes for AI-output detection tools, ideally developed in coordination with the European Telecommunications Standards Institute (ETSI) or other relevant bodies.

[375] See Katherine Trendacosta, Unfiltered: How YouTube's Content ID Discourages Fair Use and Dictates What We See Online, Electronic Frontier Foundation, 10 December 2020, available at: https://www.eff.org/wp/unfiltered-how-youtubes-content-id-discourages-fair-use-and-dictates-what-we-see-online

Table 18: Three-Pillar Check

Pillar	Status	Why?
Epistemic	🟢	Enables visibility via watermarking, fingerprinting, opt-out tags, and audit-friendly registries
Normative	🔴	Provides traceability but lacks mechanisms to assign rights or ensure payment
Systemic	🟡	Leverages trade policy and platform compliance, but lacks dedicated enforcement bodies

Note: Combine with 4.3 remuneration to close the Normative gap.

4.6. Foster collaborative governance and legal coherence

A) Promote a balanced regulatory narrative

The complexity of regulating generative AI at the intersection of innovation, intellectual property, and cultural policy calls for a strategic approach that goes beyond sector-specific interventions. While the preceding sections have addressed targeted policy actions on training data, authorship, safeguards, and transparency, their implementation ultimately depends on a **sustained, cross-sectoral dialogue** that fosters mutual understanding between AI developers, content creators, legal experts, and public authorities.

A core insight emerging from this study is that the real bottleneck in enabling fair and lawful AI training is not the absence of licensing infrastructure. In fact, robust systems already exist—particularly within Europe—that could facilitate scalable rights clearance, including through collective management organisations (CMOs). However, some actors strategically seek to avoid licensing obligations by exploiting interpretative ambiguities in the law, disputing the validity of opt-out mechanisms on formalistic grounds, or engaging in forum shopping to operate under permissive legal regimes. This behaviour is not a symptom of regulatory failure but a deliberate choice to bypass creators' rights. If unaddressed, it risks undermining trust in both the copyright and AI governance frameworks. This erosion of trust is further exacerbated by structural asymmetries in global content distribution, where U.S.-based firms not only develop the most advanced AI models but also dominate the platforms through which creative content is disseminated. Without targeted measures to strengthen the position of EU creators and cultural intermediaries, copyright reforms risk reinforcing this dependency and undermining Europe's long-term cultural and technological sovereignty.

In this context, the EU should recognise and actively support the pivotal role of CMOs in facilitating a viable and inclusive licensing ecosystem. CMOs possess the legal mandate, operational infrastructure, and stakeholder legitimacy to manage collective rights efficiently and equitably across sectors. Strengthening their role—particularly by promoting cross-border interoperability and enhancing their capacity to handle AI-related use cases—will be key to building a licensing system that is both technically scalable and normatively robust.

The EU should adopt a **balanced and forward-looking regulatory narrative** that avoids polarisation between innovation and protection. The current discourse too often presents a binary choice between unrestrained AI development and rigid copyright enforcement. Both extremes are counterproductive. A regulatory model that is "**all for innovation**" risks enabling extractive practices that erode the value of human creativity and undermine the legitimacy of AI systems. Conversely, an **overly protectionist or "copyright maximalist" stance** could stifle the development of lawful and socially beneficial AI applications. The recommendations in this paper seek to chart a **middle path**: one that facilitates legitimate and responsible access to creative works through lawful exceptions and licensing mechanisms, while ensuring that creators retain agency over their work and share in the benefits of AI-driven innovation. This balanced messaging should be explicitly embedded in future EU communications, legislative proposals, and international engagements.

B) Expand access to lawful training datasets

The EU should actively promote the **availability of high-quality, lawful datasets for AI training**, particularly by unlocking public sector content and expanding open cultural data initiatives. A strategic investment in legal datasets—whose rights are cleared or that belong to the public domain—can reduce reliance on infringing or questionable sources, especially in the early stages of model development. This could include expanding access to and the technical usability of collections hosted by initiatives such as **Europeana**, encouraging the curation of **AI-ready, rights-cleared training datasets**, and supporting the **standardisation of open licenses and metadata protocols** that clarify reusability conditions. These measures would simultaneously support AI innovation, reduce legal exposure, and relieve pressure on the use of protected works.

C) Provide tailored guidance for creators and developers

The EU should invest in the development and dissemination of **practical guidance and educational tools tailored to the needs of different stakeholders**. The legal implications of AI training and deployment remain opaque to many creators, developers, and SMEs. To address this, the Commission, in coordination with the **EUIPO and national IP offices**, should produce sector-specific guidelines such as:

- A guide for creators: "What to do if your work has been used to train AI", explaining rights, opt-out procedures, and remedies;

- A guide for developers: "How to use copyrighted content responsibly in AI training", explaining the scope of exceptions, importance of rights clearance, and best practices for dataset sourcing.

These instruments would not only promote **voluntary compliance** and transparency, but also **reduce the risk of unintentional infringement** and enhance overall confidence in the regulatory system.

D) Reinforce procedural safeguards and market-based incentives

Even if the current legal framework under Article 4 of the CDSM Directive is, in the view of this study, unsuited to govern large-scale AI training, **compliance with opt-out obligations could be strengthened through procedural reinforcement mechanisms**. In particular, **general-purpose AI providers could be required to certify**—as part of their internal governance or external transparency

documentation under **Article 53 of the AI Act**—that they have consulted EU-wide opt-out registries prior to model training. This would establish a concrete point of intersection between the copyright and AI regulatory frameworks and enhance legal accountability.

To further reinforce this nexus, the EU should explore **market access-based incentives**. Given the scale of the EU's digital market and its influence over international AI governance, transparency compliance—including respect for opt-out mechanisms and dataset disclosure—could be considered as a **precondition for entry or operation in the EU market**. Such an approach, mirroring mechanisms in the GDPR and DSA, would align enforcement with economic incentives and promote **extraterritorial compliance** by non-EU actors.

While procedural enhancements and economic incentives are necessary, they must be embedded within a broader and enforceable legal framework. In this regard, it is important to recall that the EU's existing copyright enforcement regime already includes the 2004 Enforcement Directive, which provides essential procedural instruments such as injunctions, evidentiary measures, and damages.[376] However, this Directive was not conceived with the systemic opacity and industrial scale of generative AI training in mind. Its current tools may fall short in addressing the unique enforcement challenges posed by AI, particularly when training is conducted by non-EU providers or through decentralized models. Targeted interpretive guidance or even legislative updates may be required to adapt the Directive's principles to emerging realities.

Finally, particular attention should be paid to the potential conflict between transparency obligations introduced by the AI Act and the protection of confidential business information guaranteed under the Trade Secrets Directive (Directive (EU) 2016/943). In the absence of clear procedural guidance, this tension may create legal uncertainty for GPAI providers and risk undermining the enforceability of Article 53 disclosures.

E) Address ethical risks of AI-generated content

Beyond the legal and economic dimensions, the ethical implications of AI-generated content warrant urgent consideration. Generative AI systems are increasingly used to produce synthetic media that may distort public discourse, misrepresent individuals, or perpetuate cultural and social harms. These risks, while not always covered by copyright law, intersect with EU values of human dignity, non-discrimination, and pluralism.

To address this challenge, the EU should promote ethics-by-design principles and AI content impact assessments in high-impact domains such as journalism, education, and public communication. These assessments, aligned with Article 29 of the AI Act, should include an ethical risk layer evaluating potential manipulation, misrepresentation, or discriminatory output.

At the same time, the EU could issue sector-specific ethical guidance—via the AI Office or EDMO—focused on generative content and drawing from frameworks such as the UNESCO Recommendation on the Ethics of AI. Public broadcasters, cultural institutions, and EU-funded media projects could be

[376] Directive 2004/48/EC of the European Parliament and of the Council of 29 April 2004 on the Enforcement of Intellectual Property Rights, 2004 O.J. (L 157), 45–86.

required to adopt AI content charters, specifying their standards for transparency, authorship attribution, and editorial responsibility.

Finally, the Parliament should explore the creation of a European Ethical Observatory on Generative Media, building on or linked to EDMO, to monitor evolving risks, disseminate soft-law recommendations, and support normative alignment across sectors. Embedding ethical review in both AI governance and copyright frameworks would enhance the legitimacy and social acceptability of generative content systems.

F) Address market concentration and data access asymmetries

The generative AI landscape is currently characterised by high levels of market concentration in both technical capacity and informational capital. A handful of large technology firms command privileged access to key inputs—such as high-quality copyrighted datasets, large-scale compute infrastructure, and vertically integrated deployment channels—creating structural barriers to entry for smaller developers, academic institutions, and independent creators.[377]

This dynamic raises significant competition law and market fairness concerns, particularly where exclusive or opaque dataset acquisition strategies effectively reinforce the dominance of a few actors and marginalise alternative innovation pathways. While the Digital Markets Act (DMA) already establishes obligations for gatekeepers in the digital ecosystem, its application to generative AI remains nascent and should be expanded to include training data governance and AI-as-a-service markets.
To promote greater diversity and decentralisation in generative AI development, the EU should consider the following measures:

1) Encourage the development of federated or decentralised training models, which allow multiple actors to collaboratively train models without centralising data access;

2) Support AI data commons and EU-curated repositories of legally cleared or public domain works, particularly through the expansion of initiatives like Europeana or EU-funded infrastructure under Horizon Europe;

3) Promote non-discriminatory access to essential compute and cloud services, particularly for academic and non-profit developers, possibly through state-aid frameworks or inclusion in public tenders;

4) Mandate transparency in large-scale dataset acquisition by dominant firms, with regulatory oversight of potentially exclusionary practices;

5) Integrate competition and copyright oversight in merger reviews involving large AI firms or training dataset aggregators, especially where content assets are bundled with deployment monopolies.

[377] See EUIPO, The Development of Generative Artificial Intelligence from a Copyright Perspective (May 2025), cit. at 259–262 (noting that fragmented opt-out tools, lack of harmonised standards, and uneven technical capacity place disproportionate compliance burdens on small creators and non-commercial actors).

These measures would not only foster greater market plurality and innovation in the AI ecosystem but also ensure that access to European cultural and creative assets does not become the exclusive domain of a few globally dominant actors.

G) Address fragmentation risks linked to moral rights

A distinct but often overlooked issue is the fragmented status of moral rights across Member States. Moral rights—such as the right of attribution and the right of integrity—are only partially harmonised under EU law, yet they are increasingly invoked in stakeholder debates as tools to oppose the unauthorised use of creative works for AI training. As noted in this study, a recent EU consultation found that 67% of stakeholders support allowing rightsholders to invoke moral rights even where economic rights exceptions (e.g., Article 4 CDSM) apply—particularly where AI systems mimic style, distort authorial intent, or affect professional reputation.[378] The right of integrity, in particular, was cited as potentially infringed when AI outputs reproduce an author's distinctive style or introduce distortions that could damage their public image. As generative models become more expressive, the lack of harmonisation in moral rights protection may give rise to litigation, forum shopping, or regulatory divergence. This fragmentation therefore poses both legal and political risks. Accordingly, the Commission should consider launching a review or expert consultation on the feasibility of minimum harmonisation of moral rights in the digital context, with particular focus on generative AI systems—aiming to reduce legal fragmentation while respecting national traditions.

H) Strengthen international alignment

Given the global nature of AI development and the transnational use of copyrighted content, the EU should actively promote international coordination on standards for transparency, licensing, and fair remuneration mechanisms. Engagement in multilateral fora such as WIPO, OECD, and WTO is, of course, essential to avoid fragmented regulatory approaches and ensure that European rules are interoperable with frameworks in jurisdictions like the United States, the United Kingdom, Japan, and others. Policy measures proposed at the EU level—such as dataset transparency requirements, opt-out standardisation, and collective licensing models—should be designed with international compatibility in mind. The EU should aim to lead a coalition of jurisdictions committed to protecting authors' rights while promoting responsible AI innovation, thereby reinforcing its role as a normative global power in digital governance.

Table 19: Three-Pillar Check

Pillar	Status	Why?
Epistemic	🟡	Promotes awareness through guidance tools, dataset initiatives, and ethical assessments, but lacks enforceable transparency or audit obligations.

[378] See European Commission, Study on Copyright and New Technologies: Copyright Data Management and Artificial Intelligence, cit. at 230.

Pillar	Status	Why?
Normative	🟡	Reinforces legitimate reuse frameworks and CMO licensing legitimacy, while encouraging moral rights harmonisation, but offers no binding remuneration or user rights.
Systemic	🔴	Relies on voluntary measures and strategic alignment, without proposing dedicated oversight bodies or enforceable procedural mechanisms.

Note: This section complements structural reforms by fostering the legal and ethical conditions for cross-sector alignment, but systemic accountability must be ensured through the institutional proposals in §4.1.

4.7. Conclusion

Generative AI represents a transformative technological shift—one that will continue to expand in scope, sophistication, and societal impact. Its capacity to generate text, images, music, and other creative outputs at scale challenges traditional legal categories and places increasing pressure on existing intellectual property frameworks. EU copyright law, in particular, is now being tested on multiple fronts: from the legality of using protected content as training data, to the attribution of authorship in hybrid human–AI creations, to the enforcement of rights in an environment defined by synthetic outputs and algorithmic opacity.

Yet these challenges are not insurmountable. As this paper has argued, the European Union is well positioned to respond—not by overhauling the copyright acquis, but through **adaptive governance**: refining the application of existing instruments such as the CDSM Directive, aligning implementation with emerging frameworks like the AI Act, and filling critical gaps through targeted interventions. This study also highlights two underlying structural risks that demand attention: first, the **erosion of fair bargaining power for authors** in negotiations over AI training uses; second, the **displacement of human creativity** through the mass deployment of generative content across digital platforms. These dynamics expose fundamental weaknesses in the current market design—where rights are often unenforceable in practice, and visibility in distribution is increasingly determined by algorithmic amplification. Addressing these failures is not only a matter of fairness but essential for preserving the diversity, sustainability, and long-term viability of Europe's creative economy.

In order to counter these risks effectively, the EU must use these tools in tandem and with foresight, the EU can establish a legal and ethical ecosystem in which **AI innovation can flourish without undermining the creative economy** that fuels Europe's cultural diversity, democratic discourse, and knowledge systems.

The recommendations advanced in this paper seek to future-proof the legal framework in four key ways:

1) **Closing regulatory gaps**, particularly around transparency, remuneration, and traceability;

2) **Clarifying normative boundaries**, including authorship standards, liability attribution, and the distinction between data analysis and content reproduction;

3) **Reinforcing safeguards and procedural protections**, through interpretative guidance, technical standards, and interoperable disclosure mechanisms;

4) And **fostering inclusive governance**, through structured dialogue, educational resources, and investment in lawful training datasets.

These proposals do not treat innovation and authorship as opposing values. Instead, they articulate a **balanced regulatory model**—one that enables responsible AI development, ensures respect for human creativity, and reinforces Europe's dual leadership in technological advancement and cultural production.

The governance of generative AI will shape not only future markets, but also the ways in which knowledge, culture, and meaning are produced and shared. The EU has an opportunity—and arguably a responsibility—to lead by example, demonstrating that digital transformation can be steered toward **inclusive, sustainable, and rights-respecting outcomes**. A timely and coordinated policy response is warranted to ensure that AI innovation aligns with Europe's legal traditions and creative values.

Ensuring the continuity and effectiveness of the proposed reforms requires a coordinated institutional response. In the immediate term, a **High-Level Expert Group (HLEG)** could be tasked with developing enforceable technical standards, piloting remuneration mechanisms, and assessing the feasibility of machine-readable opt-out solutions. In parallel, the **JURI Committee may wish to establish a dedicated Working Group on AI and Copyright**, functioning as a parliamentary platform to oversee the HLEG's output, facilitate legislative follow-up, and promote structured engagement with other committees and stakeholders. This dual mechanism would help bridge expert analysis and political oversight, reinforcing the Parliament's central role in shaping a coherent and future-oriented copyright framework. Looking ahead, a permanent **AI & Copyright Unit** embedded within the EU AI Office could institutionalise these efforts, ensuring long-term policy alignment, audit capacity, and regulatory continuity. Taken together, these mechanisms form a phased and complementary governance structure, each serving a distinct role in the transition from experimentation to implementation.

In order to complement the preceding legal and policy analysis, this study outlines three plausible futures for Europe's creative sectors depending on the level of regulatory intervention adopted by 2030. These are not predictions but illustrative trajectories: one aligned with full implementation of this study's recommendations, one reflecting partial uptake, and one assuming continued inaction.

- **Guided Progress (Optimistic):** full uptake of recommendations leads to legal certainty, remuneration, and robust EU participation in foundation model development.

- **Litigious Status Quo (Intermediate)**: partial or fragmented implementation yields case-by-case rulings, weak incentives, and market marginalisation.

- **Creative Erosion (Regressive)**: regulatory inaction leads to unchecked AI use, market extraction, and collapse of sustainable creative industries.

The variables assessed include legal frameworks for AI training, dataset transparency, market share of EU-developed models, income trends in the creative sector, and cultural-linguistic diversity.[379]

Table 20: Scenario Outlook 2030: Strategic Futures for EU Copyright Governance

Key variable	🟢 **Guided Progress** (full uptake)	🟡 **Litigious Status Quo** (partial uptake)	🔴 **Creative Erosion** (no action)
Legal basis for AI training	Opt-in framework + EU-wide collective licence	Court rulings and opt-out-based exceptions	No licensing; opt-out ineffective or ignored
Dataset transparency	EUIPO registry + AI Office audits; public dataset logs	Voluntary, Member-State-level disclosures	No access to training logs; full model opacity
Market share of EU-built foundation models	≈ 25 % of general-purpose model market	< 10 %; US giants dominate	≈ 0 %; Europe a pure consumer market
Economic health of creative sectors	Rights income up ≈ +15 % vs 2023; SMEs participate	Flat growth; revenue captured by a few majors	Median creator income down ≈ –40 %
Cultural-linguistic diversity	Multilingual AI output; minority languages visible	English-heavy output; EU content marginal	Global narrative homogenised; loss of local voices

These risk scenarios underscore the strategic choices facing the European Union—not only in shaping its internal copyright regime but also in defining its position in the global digital order. A regulatory framework grounded in **transparency, fair remuneration, and systemic accountability** can unlock sustainable innovation and cultural pluralism, while reinforcing the EU's leadership in **normative AI governance**. By contrast, inaction risks the long-term erosion of Europe's **creative economy, legal coherence, and digital sovereignty**. In a global environment where major jurisdictions may opt for minimal or no regulation, the EU could find itself uniquely constrained—its cultural assets exposed to extraction, and its markets transformed into permissive training grounds.

As a concrete decision-support tool, the table below applies the Three-Pillar Accountability Test to each of the key recommendations discussed in Section 4. The table below applies the three-pillar grid to every recommendation illustrated in this chapter. Measures marked 🟢 in all three columns are necessary cornerstones of a balanced EU solution, while amber or red cells flag the residual gaps the Parliament may wish to close in trilogue.

[379] The figures indicated in the table are indicative and serve only to illustrate the potential magnitude of impact under each scenario. They are not intended as precise forecasts but as directional outcomes, grounded in current policy and market trends.

Table 21: Three-Pillar Check

Measure (short label)	Epistemic	Normative	Systemic	Explanation
4.1 Governance & Enforcement	🟡	🟡	🟡	Proposes structural reforms, but lacks immediate legal effect or enforcement powers.
4.2 TDM Fix (Clarify Art. 4 / Opt-in)	🟡	🟡	🔴	Improves legal clarity and opt-out visibility, but lacks auditing and enforceability mechanisms.
4.3 Remuneration Mechanisms	🟡	🟢	🟡	Establishes fair compensation rights, but depends on future metadata infrastructure and institutional oversight.
4.4 Authorship & Protection Status	🔴	🟢	🟡	Affirms the public domain status of AI-only outputs, but lacks traceability and binding safeguards.
4.5 Safeguards & Traceability Tools	🟢	🔴	🟡	Enables traceability and transparency, but does not establish rights or compensation mechanisms.

Reading guide: Cells turn green only when the measure fully addresses the relevant accountability pillar. A complete legislative package must therefore combine at least 4.2 (TDM Fix – Clarify Art. 4 / Opt-in) + 4.3 (Remuneration Mechanisms) + 4.5 (Safeguards & Traceability Tools) — or equivalent — to achieve 🟢 across the board.

Based on this assessment, **priority should be given** to implementing the TDM fix (Clarify Article 4 / Opt-in) (4.2), the remuneration mechanisms (4.3), and the safeguards and traceability tools (4.5), as these measures collectively address the most critical gaps across all three accountability dimensions. Without this triad—ensuring legal clarity, fair compensation, and verifiable transparency—neither legal coherence nor sustainable innovation can be achieved.

REFERENCES

- Abbott, Ryan Benjamin and Rothman, Elizabeth, Disrupting Creativity: Copyright Law in the Age of Generative Artificial Intelligence, 75 Florida Law Review 1141 (2023)
- Agrawal, Ajay et al., Prediction Machines: The Simple Economics of Artificial Intelligence (Harv. Bus. Rev. Press 2018)
- Asperti, Andrea and Tonelli, Valerio, Comparing the latent space of generative models, 35 Neural Computing and Applications 3155–3172 (2023)
- Balkin, Jack, The Path of Robotics Law, 6 Calif. L. Rev. 45 (2015)
- Barfield, Woodrow & Pagallo, Ugo (eds), Research Handbook on the Law of Artificial Intelligence (Edward Elgar 2018)
- Baumann, Malte, Generative KI und Urheberrecht – Urheber und Anwender im Spannungsfeld, NJW - Neue Juristische Wochenschrift, 3673-3678 (2023)
- Bender, Emily M. and Koller, Alexander, Climbing Towards NLU: On Meaning, Form, and Understanding in the Age of Data, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020): 5185–5198
- Bengio, Yoshua; Lecun, Yann; Hinton, Geoffrey, Deep learning for AI, 64 Communications of the ACM 58-65 (2021)
- Blaszczyk, Matt, Impossibility of Emergent Works' Protection in U.S. and EU Copyright Law, 25 North Carolina Journal of Law & Technology 1 (2023)
- Bommasani, Rishi et al., On the Opportunities and Risks of Foundation Models, Preprint at https://doi.org/10.48550/arXiv.2108.07258 (2022)
- Bonadio, Enrico et al., Can artificial intelligence infringe copyright? Some reflections. Research Handbook on Intellectual Property and Artificial Intelligence (Ryan Abbott ed.) (Edward Elgar 2022)
- Bonadio Enrico, Lucchi Nicola, Mazziotti, Giuseppe, Will Technology-Aided Creativity Force Us to Rethink Copyright's Fundamentals? Highlights from the Platform Economy and Artificial Intelligence, 53(8) IIC - International Review of Intellectual Property and Competition Law 1174 (2022)
- Bonadio, Enrico & McDonagh, Luke, Artificial Intelligence as Producer and Consumer of Copyright Works: Evaluating the Consequences of Algorithmic Creativity, Intellectual Property Quarterly 2, 112-137 (2020)
- Bonadio, Enrico et al., 'Intellectual property aspects of robotics' European Journal of Risk Regulation 9(4) 655–676 (2018)
- Bonadio, Enrico & Nicola Lucchi, 'How Far Can Copyright Be Stretched? Framing the Debate on Whether New and Different Forms of Creativity Can Be Protected', Intellectual Property Quarterly, 115-135. (2019)

- Boyden, Bruce E., Emergent Works, 39 Colum. J.L. & Arts 377, (2016)
- Brauneis, Bob, Copyright and the Training of Human Authors and Generative Machines, 48 Columbia Journal of Law and the Arts 1 (2025)
- Bridy, Annemarie, The Evolution of Authorship: Work Made by Code, 39 Colum. J.L. & Arts 395 (2016)
- Buccafusco, Christopher, A Theory of Copyright Authorship, 102 Virginia Law Review 1229-1295 (2016)
- Buick, Adam, Copyright and AI training data—transparency to the rescue? 20 Journal of Intellectual Property Law & Practice, 182 -192 (2025)
- Colangelo, Giuseppe, A Competition Policy Analysis of Copyright Protection in Gen AI, Singapore Journal of Legal Studies, forthcoming (2025). Available at https://ssrn.com/abstract=5201510
- Calo, Ryan, et al. (eds). Robot Law, (Edward Elgar 2016)
- Chang, Chien-Yi and He, Xin, The Liabilities of Robots.Txt. University of Hong Kong Faculty of Law Research Paper No. 2025/06. Available at SSRN: https://ssrn.com/abstract=5159436
- De Cock Buning, M., Autonomous Intelligent Systems as Creative Agents under the EU Framework for Intellectual Property, 7 Eur. J. Risk Reg. 310 (2016)
- De Cremer, David et al., How Generative AI Could Disrupt Creative Work, Harvard Business Review (Apr. 13, 2023). Available at https://hbr-org.sare.upf.edu/2023/04/how-generative-ai-could-disrupt-creative-work
- Denicola, Robert C., Ex Machina: Copyright Protection for Computer-Generated Works, 69 Rutgers U. L. Rev. 251 (2016)
- Dermawan, Artha, Text and Data Mining Exceptions in the Development of Generative AI Models: What the EU Member States Could Learn from the Japanese "Nonenjoyment" Purposes, 27 J. World Intell. Prop. 44 (2023)
- Dermawan, Artha and Mezei, Péter, Artificial Intelligence and Consensus-Based Remuneration Regime in Southeast Asia (November 7, 2023). Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4625850
- Desai, Deven R. & Riedl, Mar, Between Copyright and Computer Science: The Law and Ethics of Generative AI, 22 Northwestern Journal of Technology & Intellectual Property 55-108 (2024)
- Dornis, Tim W., Of "authorless works" and "inventions without inventor" - the muddy waters of "AI autonomy" in intellectual property doctrine, 43 European Intellectual Property Review 570 (2021)
- Dornis, Tim W., Artificial Creativity: Emergent Works and the Void in Current Copyright Doctrine, 22 Yale J. L. & Tech. 1 (2020)
- Dornis, Tim W., Wigmorian Copyright: Law, Economics, and Socio-Cultural Evolution," Intellectual Property Quarterly (IPQ), 3/2018, 159-180

- Dornis, Tim W., Artificial Creativity: Emergent Works and the Void in Current IP Law," 22 Yale Journal of Law & Technology (Yale J. L. & Tech.) 2020, pp. 1-60
- Dornis, Tim W. and Stober, Sebastian, Urheberrecht und Training generativer KI-Modelle - Technologische und juristische Grundlagen) NOMOS Verlag (Baden-Baden 2024)
- Dornis, Tim, The Training of Generative AI Is Not Text and Data Mining, 47 European Intellectual Property Review 65-78, 2025
- Dornis, Tim Generative AI, Reproductions Inside the Model, and the Making Available to the Public. IIC - International Review of Intellectual Property and Competition Law (2025)
- Ducato, Rossana and Strowel, Alain 'Ensuring Text and Data Mining: Remaining Issues with the EU Copyright Exceptions and Possible Ways Out' 43 European Intellectual Property Review 322–337 (2021)
- Durante, Zane et al., Agent AI: Surveying the Horizons of Multimodal Interaction (arXiv:2401.03568v2) (2024). Available at https://arxiv.org/abs/2401.03568
- Dusollier, Severine et al., Copyright and Generative AI: Opinion, 16 JIPITEC 121 (2025)
- Durantaye, Katharina de la, Control and Compensation. A Comparative Analysis of Copyright Exceptions for Training Generative AI, IIC - International Review of Intellectual Property and Competition Law 1-34 (2025)
- European Commission: Directorate-General for Communications Networks, Content and Technology, Hartmann, C. et al., Trends and developments in artificial intelligence – Challenges to the intellectual property rights framework – Final report, Publications Office of the European Union, 2020. Available at https://data.europa.eu/doi/10.2759/683128
- European Commission: Study on copyright and new technologies: Copyright data management and artificial intelligence, Publications Office of the European Union 2022. Available at https://op.europa.eu/publication-detail/-/publication/cc293085-a4da-11ec-83e1-01aa75ed71a1
- European Union Intellectual Property Office, Study on the Impact of Artificial Intelligence on the Infringement and Enforcement of Copyright and Designs report" (EUIPO, 2022). Available at https://www.euipo.europa.eu/en/publications/study-on-the-impact-of-artificial-intelligence-on-the-infringement-and-enforcement-of-copyright-and-designs
- European Union Intellectual Property Office, Development of Generative Artificial Intelligence from a Copyright Perspective (May 2025). Available at https://www.euipo.europa.eu/en/publications/genai-from-a-copyright-perspective-2025
- European Parliament, European Parliament resolution of 20 October 2020 on intellectual property rights for the development of artificial intelligence technologies P9_TA(2020)0277. Available at https://www.europarl.europa.eu/doceo/document/TA-9-2020-0277_EN.html
- Fayyad, Usama et al., From Data Mining to Knowledge Discovery in Databases, 17 AI Magazine 37–54 (1996)

- Fernández Carballo-Calero, Pablo, La propiedad intelectual de las obras creadas por inteligencia artificial (Aranzadi 2021)
- Floridi, Luciano, The European Legislation on AI: a Brief Analysis of its Philosophical Approach, 34 Philosophy & Technology 215 (2021)
- Floridi, Luciano, AI as Agency without Intelligence: on Chat GPT, Large Language Models and Other Generative Models, 36 Philosophy & Technology 1 (2023)
- Fong Terrance et al., A Survey of Socially Interactive Robots, 42 Robotics & Autonomous Systems 143 (2003)
- Franceschelli, Giorgio and Musolesi, Mirco, "Copyright in generative deep learning" 4 Data & Policy e17 (2022)
- Franceschelli, Giorgio and Musolesi, Mirco, On the creativity of large language models. AI & Soc 1-11 (2024)
- Frosio, Giancarlo Four theories in search of an A(I)uthor, in Ryan Abbott (ed), Handbook of Artificial Intelligence and Intellectual Property 156-178 (Edward Elgar 2022)
- Frosio, Giancarlo "Should We Ban Generative AI, Incentivise It or Make It a Medium for Inclusive Creativity?" in E Bonadio and C Sganga (eds), A Research Agenda for EU Copyright Law 61 (Cheltenham, Edward Elgar, 2025)
- Geiger, Christophe et al., Text and Data Mining in the Proposed Copyright Reform: Making the EU Ready for an Age of Big Data?, 49 IIC - International Review of Intellectual Property and Competition Law 49, 814–844 (2018)
- Geiger, Christophe, Building an Ethical Framework for Intellectual Property in the EU: Time to Revise the Charter of Fundamental Rights, in G. Ghidini and V. Falce (eds), Reforming Intellectual Property 77 (Edward Elgar, 2022)
- Geiger, Christophe and Iaia, Vincenzo, The forgotten creator: towards a statutory remuneration right for machine learning of generative AI. 52 Computer Law & Security Review 1-9 (2024)
- Geiger, Christophe, Elaborating a Human Rights Friendly Copyright Framework for Generative AI, 55 International Review of Intellectual Property and Competition Law 1129–1165 (2024)
- Gervais, Daniel J., The Machine as Author, 105 Iowa L. Rev. 2053 (2020)
- Gervais, Daniel J, The Human Cause, in R Abbott (ed.), Research Handbook on Intellectual Property and Artificial Intelligence (Cheltenham, Edward Elgar 2022)
- Gervais, Daniel, Towards an effective transnational regulation of AI, 38 AI & Society 391 (2023)
- Ginsburg, Jane, People Not Machines: Authorship and What It Means in the Berne Convention, 49 IIC - International Review of Intellectual Property and Competition Law 131 (2018)
- Ginsburg, Jane and Budiardjo, Luke, Authors and Machines, 34 Berkeley Technology Law Journal 343 (2019)

- Gliściński, Konrad, Polish Implementation of TDM Exceptions – General Characteristics, Stockholm IP Law Review 2024#2, 9-18 (April 2025)
- Goldberg, Yoav. A Primer on Neural Network Models for Natural Language Processing. 57 Journal of Artificial Intelligence Research, 345–420 (2016)
- Goldberg, Yoav, Neural Network Methods for Natural Language Processing (Cham, Springer 2017)
- Goldstein, Caroline, Rembrandt's Revered 'Night Watch' Was Cut Up to Fit Through a Door. With A.I., You Can See It Whole for the First Time in 300 Years, ARTNET NEWS (June 23, 2021), https://news.artnet.com/art-world/operation-night-watch-1982686
- Goodfellow, Ian J. et al., Generative Adversarial Nets. 2 Proceedings of the 27th International Conference on Neural Information Processing Systems 2672-2680 (2014)
- Grimmelmann, James, There's No Such Thing as a Computer-Authored Work - And It's a Good Thing, Too, 39 Colum. J.L. & Arts 403 (2016)
- Guadamuz, Andres, A Scanner Darkly: Copyright Liability and Exceptions in Artificial Intelligence Inputs and Outputs, 73 GRUR Int. 111-127 (2024)
- Guadamuz, Andres, Do androids dream of electric copyright? Comparative analysis of originality in artificial intelligence generated works, Intellectual Property Quarterly, 169 (2017)
- Guadamuz, Andres, Artificial intelligence and copyright, WIPO Magazine (2017)
- Hamann, Hanjo, Artificial Intelligence and the Law of Machine-ReadabilityA Review of Human-to-Machine Communication Protocols and their (In)Compatibility with Article 4(3) of the Copyright DSM Directive, 15 JIPITEC 102-121 (2024)
- Holzmüller, Tobias, Gesellschaft für musikalische Aufführungs-und mechanische Vervielfältigungsrechte (GEMA), personal communication, email to author, (May 2025)
- Hilgendorf, Eric & Seidel, Uwe, Robotics, Autonomics, and the Law (Nomos 2017)
- Hine, Emmie and Floridi, Luciano, Artificial Intelligence with American Values and Chinese Characteristics: A Comparative Analysis of American and Chinese Governmental AI Policies (January 11, 2022). Available at SSRN: https://ssrn.com/abstract=4006332
- Hoc, Jean-Michel, From human-machine interaction to human-machine cooperation, 43 Hergonomics 833 (2000)
- Hugenholtz, P. Bernt (ed.), The Future of Copyright in a Digital Environment (Kluwer 1996)
- Hugenholtz, P. Bernt and Quintais, Joao Pedro, Copyright and Artificial Creation: Does EU Copyright Law Protect AI-Assisted Output?, 52 IIC - International Review of Intellectual Property and Competition Law, 1190 (2021)
- Hughes, Justin, The Philosophy of Intellectual Property, 77 Geo. L.J. 287 (1988)
- Hughes, Justin, A Short History of "Intellectual Property" in Relation to Copyright, 33 Cardozo L. Rev. 1293 (2012)

- Judge, Elizabeth F. and Gervais, Daniel, Of Silos and Constellations: Comparing Notions of Originality in Copyright Law, 27 Cardozo Arts & Entertainment Law Journal 375 (2009)
- Kaminski, Margot E., Authorship, Disrupted, 51 U.C. Davis L. Rev. 589 (2017)
- Kang, Hyunjin and Lou, Chen, AI agency vs. human agency: understanding human–AI interactions on TikTok and their implications for user engagement, 27 Journal of Computer-Mediated Communication 1-13 (2022)
- Keisner A. et al., Robotics: Breakthrough Technologies, Innovation, Intellectual Property'. Foresight and STI Governance 10 (2): 7–27 (2016)
- Keisner A. et al., Breakthrough Technologies – Robotics and IP, 6 WIPO Magazine (2016)
- Khoury, Amir. H. Intellectual Property Rights for Hubots: On the Legal Implications of Human-like Robots as Innovators and Creators, 35 Cardozo Arts & Ent LJ 635 (2017)
- Kretschmer, Martin, Private Copying and Fair Compensation: An Empirical Study of Copyright Levies in Europe - A Report for the UK Intellectual Property Office (2011). Available at https://ssrn.com/abstract=2710611
- Kretschmer, Martin et al., Artificial Intelligence and Intellectual Property: Copyright and Patents—A Response by the CREATe Centre to the UK Intellectual Property Office's Open Consultation, 17 Journal of Intellectual Property Law & Practice 321-326 (2022)
- Kretschmer, Martin et al., Copyright Law and the Lifecycle of Machine Learning Models, 55 IIC - International Review of Intellectual Property and Competition Law 110–138 (2024)
- Iaia, Vincenzo, To Be, or Not to Be ... Original Under Copyright Law, That Is (One of) the Main Questions Concerning AI-Produced Works, 71 GRUR International, 793–812 (2022)
- Lauber-Rönsberg, Anne and Hetmank, Sven, The concept of authorship and inventorship under pressure: Does artificial intelligence shift paradigms?, 14 Journal of Intellectual Property Law & Practice, 570 (2019)
- LeCun, Yann, Bengio, Yoshua & Hinton, Geoffrey, Deep learning. 521 Nature 436–444 (2015)
- Lee, Katherine et al, Talkin' 'Bout AI Generation: Copyright and the Generative-AI Supply Chain, 71 Journal of the Copyright Society of the U.S.A. (forthcoming 2024)
- Lemley, Mark A. & Casey, Bryan, Remedies for Robots, 86 U. Chi. L. Rev. 1311 (2019)
- Lemley, Mark A. & Casey, Bryan, Fair Learning, 99 Texas Law Review 743 (2021)
- Lim, Daryl, AI & IP: Innovation & Creativity in an Age of Accelerated Change, 52 Akron L. Rev. 813 (2018)
- Lim, Daryl, Generative ai and copyright: principles, priorities and practicalities, 18 Journal of Intellectual Property Law & Practice 841 (2023)
- Longpre, Shayne et al. A large-scale audit of dataset licensing and attribution, 6 Nature Machine Intelligence 975–987 (2024)

- Lucchi, Nicola & Bonadio, Enrico (eds.), Non-Conventional Copyright: Do New and Non Traditional Works Deserve Protection? (Edward Elgar, 2018)
- Lucchi, Nicola & Laukyte, Migle, Creative AI: The Complex Relationship between Human Inventiveness and Intellectual Property, in BioLaw Journal Vol. 22 (3) pp.169-183 (2022)
- Lucchi, Nicola, ChatGPT: A Case Study on Copyright Challenges for Generative Artificial Intelligence Systems. European Journal of Risk Regulation, 1-23. doi:10.1017/err.2023.59 (2023)
- Marušić, Branka TDM Exception or Limitation –Methodology of Implementation in the EU Member States: Creating Cohesion or Diversion?, Stockholm IP Law Review 2024#2, 19-24 (April 2025)
- Margoni, Thomas and Kretschmer, Martin, A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology, 71 GRUR International 685–701 (2022)
- Margoni, Thomas, TDM and generative AI: Lawful access and opt-outs. Auteurs en Media 175 - 188 (2024)
- Mazziotti, Giuseppe, What Is the Future of Creators' Rights in an Increasingly Platform-Dominated Economy?, 51 IIC-International Review of Intellectual Property and Competition 1027–1032 (2020)
- McCaffrey, Tony and Spector, Lee, An approach to human–machine collaboration in innovation. 32 AI EDAM, 1-15 (2018)
- McCutcheon, Jani, The Vanishing Author in Computer-Generated Works: A Critical Analysis of Recent Australian Case Law, Melbourne University Law Review 36 (2013)
- Mezei, Peter, A saviour or a dead end? Reservation of rights in the age of generative AI, 46 Eur. IP Rev. 461 (2024)
- Mezei, Péter, "You Ain't Seen Nothing Yet" - Arguments against the Protectability of AI-generated Outputs by Copyright Law. In: Maurizio Borghi - Roger Brownsword (eds.): Informational Rights and Informational Wrongs: A Tapestry for Our Times, 126-143 (Routledge 2023)
- Mezei, Péter, From Leonardo to the Next Rembrandt - The Need for AI-Pessimism in the Age of Algorithms, UFITA – Archiv für Medienrecht und Medienwissenschaft, 390 (2020)
- Mezei, Peter and Harkai, István, Enforcement of Copyrights over the Internet - A Review of the Recent Case Law of the CJEU, 21(4) Journal of Internet Law 1 (2017)
- Mimler, Mark, Bonadio E. et al., Implications of artificial intelligence in action - a Jamaican perspective. European Intellectual Property Review, 44(10), pp. 611–622 (2022)
- Mokhtarian, Edmund, The Bot Legal Code: Developing a Legally Compliant Artificial Intelligence, 21 Vand. J. Ent. & Tech. L. 145 (2018)
- Novelli, Claudio et al., Generative AI in EU law: Liability, privacy, intellectual property, and cybersecurity, 55 Computer Law & Security Review 1-16 (2024)

- OECD, Intellectual property issues in artificial intelligence trained on scraped data, OECD Artificial Intelligence Papers, No. 33, (OECD Publishing, Paris, 2025). Available at https://doi.org/10.1787/d5241a23-en
- Pasquale, Franck New Laws of Robotics: Defending Human Expertise (Harvard Univ. Press 2020)
- Pasquale, Frank and Sun, Haochen, Consent and Compensation: Resolving Generative AI's Copyright Crisis, 110 Virginia Law Review Online 207–47 (2024)
- Petit, Nicolas and De Cooman Jerome, Models of law and regulation for AI, Working Paper, EUI RSCAS, 2020/63
- Peukert, Alexander, et al. European Copyright Society - Comment on Copyright and the Digital Services Act Proposal, 53 IIC - International Review of Intellectual Property and Competition Law 358 (2022)
- Peukert, Alexander, Copyright in the Artificial Intelligence Act – A Primer, 73 GRUR International 497-509 (2024)
- Peukert, Christian, Copyright Levies and Cloud Storage: Ex-Ante Policy Evaluation with a Field Experiment, 53 Research Policy, 1-12 (2024)
- Picht, Peter Georg and Thouvenin, Florent AI and IP: Theory to Policy and Back Again – Policy and Research Recommendations at the Intersection of Artificial Intelligence and Intellectual Property, 54 International Review of Intellectual Property and Competition Law 916–940 (2023)
- Pukas, Jonathan, KI-Trainingsdaten und erweiterte kollektive Lizenzen: Generierung von Werken als KI-Trainingsdaten auf Basis erweiterter kollektiver Lizenzen, GRUR 614 (2023)
- Quang, Jenny "Does Training AI Violate Copyright Law?" (2021) 36 Berkeley Technology Law Journal 1407
- Quintais, João Pedro, Mezei, Péter, Harkai, István, Vieira Magalhães, João, Katzenbach, Christian, Schwemer, Sebastian Felix, Riis, Thomas: Copyright Content Moderation in the EU: An Interdisciplinary Mapping Analysis (August 1, 2022). Available at SSRN: https://ssrn.com/abstract=4210278
- Quintais, João Pedro, Generative AI, copyright and the AI Act, 56 Computer Law & Security Review 1-17 (2025)
- Ramalho, Ana Will Robots Rule the (Artistic) World? A Proposed Model for the Legal Status of Creations by Artificial Intelligence Systems, 21 Journal of Internet Law 1 (2017)
- Ramhalo, Ana, Intellectual Property Protection for AI-generated Creations (Routledge 2021)
- Razmerita, Liana et al., Collaboration in the Machine Age: Trustworthy Human-AI Collaboration. In: Virvou, M., Tsihrintzis, G.A., Jain, L.C. (eds) Advances in Selected Artificial Intelligence Areas. Learning and Analytics in Intelligent Systems, Springer 333-356 (2020)
- Ren, Minglun et al H. Human-machine Collaborative Decision-making: An Evolutionary Roadmap Based on Cognitive Intelligence.15 International Journal of Social Robotics, 15, 1101–1114 (2023)

- Riemer, Kai and Peter, Sandra, Conceptualizing generative AI as style engines: Application archetypes and implications, 79 International Journal of Information Management 1-15, (2024)
- Rosati, Eleonora, The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market – Technical Aspects, Policy Department for Citizens' Rights and Constitutional Affairs, European Parliament, (February 2018), PE 604.942. Available at https://www.europarl.europa.eu/RegData/etudes/IDAN/2018/604941/IPOL_IDA(2018)604941_EN.pdf
- Rosati, Eleonora Is text and data mining synonymous with AI training? 19 Journal of Intellectual Property Law & Practice, 851 (2024)
- Rosati, Eleonora, No Step-Free Copyright Exceptions: The Role of the Three-step in Defining Permitted Uses of Protected Content (including TDM for AI-Training Purposes), 46 European Intellectual Property Review 262-274 (2024)
- Rosati, Eleonora, No step-free copyright exceptions: the role of the three-step in defining permitted uses of protected content (including TDM for AI-training purposes), 46 European Intellectual Property Review 262-274 (2024)
- Rosati, Eleonora Infringing AI: Liability for AI-Generated Outputs under International, EU, and UK Copyright Law. European Journal of Risk Regulation. Published online 2024:1-25. doi:10.1017/err.2024.72
- Rosati, Eleonora, Copyright in the Digital Single Market: Article-by-Article Commentary to the Provisions of Directive 2019/790 (Oxford, 2021)
- Rosati, Eleonora, Copyright as an Obstacle or an Enabler? A European Perspective on Text and Data Mining and its Role in the Development of AI Creativity, 27 Asia Pacific Law Review 198–217 (2019)
- Sag, Matthew, Fairness and Fair Use in Generative AI Authors, 92 Fordham L. Rev. 1887 (2024)
- Sag, Matthew and Yu, Peter K., The Globalization of Copyright Exceptions for AI Training, 74 Emory Law Journal, (2025)
- Samuelson, Pamela, Allocating Ownership Rights in Computer-Generated Works, 47 U. Pitt. L. Rev. 1185 (1986)
- Samuelson, Pamela, Generative AI Meets Copyright 381 Science 158-159 (2023)
- Schack, Haimo Auslesen von Webseiten zu KI-Trainingszwecken als Urheberrechtsverletzung de lege lata et ferenda 77 NJW – Neue Juristische Wochenschrift 113-118 (2024)
- Senftleben, M., & Buijtelaar, L. D., Robot Creativity: An Incentive-Based Neighbouring Rights Approach. 42 European Intellectual Property Review, 797 (2020)
- Senftleben, Martin, Compliance of national TDM rules with international copyright law—an overrated nonissue? 53 IIC- International Review of Intellectual Property and Competition Law 1477 (2022)

- Senftleben, Martin Study on EU copyright and related rights and access to and reuse of data (Publications Office of the European Union, 2022, available at https://data.europa.eu/doi/10.2777/78973
- Senftleben, Martin, A Tax on Machines for the Purpose of Giving a Bounty to the Dethroned Human Author – Towards an AI Levy for the Substitution of Human Literary and Artistic Works (28 January 2022) https://ssrn.com/abstract=4123309
- Senftleben, Martin, Generative AI and Author Remuneration. 54 IIC - International Review of Intellectual Property and Competition Law, 1535–1560 (2023)
- Sherman, Brad and Wiseman, Leanne (eds), Copyright and the Challenge of the New (Kluwer Law International, 2012)
- Sobel, Benjamin, Artificial Intelligence's Fair Use Crisis, 41 Colum. J. L. & Arts 45 (2017)
- Sobel, Benjamin, "A Taxonomy of Training Data: Disentangling the Mismatched Rights, Remedies, and Rationales for Restricting Machine Learning", in R Hilty et al (eds), Artificial Intelligence and Intellectual Property (Oxford, Oxford University Press 2021)
- Stieper, Malte and Denga, Michael, The international reach of EU copyright through the AI Act, Institut für Wirtschaftsrecht 2024. Available at http://dx.doi.org/10.25673/116949
- Strowel, Alain, ChatGPT and Generative AI Tools: Theft of Intellectual Labor?, 54 IIC- International Review of Intellectual Property and Competition Law 491(2023)
- Trapova, Alina and Mezei, Péter, Robojournalism – A Copyright Study on the Use of Artificial Intelligence in the European News Industry, 71 GRUR International 589 (2022)
- Tyagi, Kalpana, Copyright, text & data mining and the innovation dimension of generative AI, 19 Journal of Intellectual Property & Practice 557–570 (2024)
- Ueno, T., The Flexible Copyright Exceptions for 'Non-Enjoyment' Purposes – Recent Amendment in Japan and Its Implication, 70 GRUR International 145–152 (2021)
- Vesala, Juha, Developing Artificial Intelligence-Based Content Creation: Are EU Copyright and Antitrust Law Fit for Purpose? 54 International Review of Intellectual Property and Competition Law 351 (2023)
- Wang, Jiachen T. et al., An Economic Solution to Copyright Challenges of Generative AI, arXiv (Apr. 2024), https://arxiv.org/abs/2404.13964
- Welser, Marcus, Generative KI und Urheberrechtsschranken, GRUR-Prax 516–520 (2023)
- Yang, S. Alex and Zhang, Angela Huyue, Generative AI and Copyright: A Dynamic Perspective (February 4, 2024). Available at SSRN: https://ssrn.com/abstract=4716233
- Yanisky-Ravid, Shlomit & Liu, Xiaoqiong, When Artificial Intelligence Systems Produce Inventions: An Alternative Model for Patent Law at the 3A Era, 39 Cardozo L. Rev. 2215 (2018)

- Yanisky-Ravid, Shlomit, Generating Rembrandt: 2017 Visionary Article in Intellectual Property Law: Generating Rembrandt: Artificial Intelligence, Copyright, and Accountability in the 3A Era–The Human Like Authors Are Already Here – A New Model, 2017 Mich. St. L. Rev. 659 (2017)
- Yasmine, Zoya Getty Images v Stability AI: Why Should UK Copyright Law Require Licences for Text and Data Mining Used to Train Commercial Generative AI Systems, 1 Cambridge Journal of Artificial Intelligence 108–120 (2024)
- Yu, Peter K., The Algorithmic Divide and Equality in the Age of Artificial Intelligence, 72 Fla. L. Rev. 331 (2020)
- Yu, Robert, The Machine Author: What Level of Copyright Protection Is Appropriate for Fully Independent Computer-Generated Works?, 165 U. Pa. L. Rev. 1245 (2017)

This study examines how generative AI challenges core principles of EU copyright law. It highlights the legal mismatch between AI training practices and current text and data mining exceptions, and the uncertain status of AI-generated content. These developments pose structural risks for the future of creativity in Europe, where a rich and diverse cultural heritage depends on the continued protection and fair remuneration of authors. The report calls for clear rules on input/output distinctions, harmonised opt-out mechanisms, transparency obligations, and equitable licensing models. To balance innovation and authors' rights, the European Parliament is expected to lead reforms that reflect the evolving realities of creativity, authorship, and machine-generated expression.

This study was commissioned by the European Parliament's Policy Department for Justice, Civil Liberties and Institutional Affairs at the request of the Committee on Legal Affairs.

PE 774.095
IUST/2025/B/JURI/IC/005
Print ISBN 978-92-848-2839-5 | doi:10.2861/0365517 | QA-01-25-130-EN-C
PDF ISBN 978-92-848-2838-8 | doi:10.2861/9120512 | QA-01-25-130-EN-N