

AI, COPYRIGHT AND DATA TRANSPARENCY

RITA MATULIONYTE*

The unauthorized use of creative content in Artificial Intelligence (AI) training has led to an outrage among copyright holders and already resulted in over forty legal actions against AI producers in the United States (US) alone. While legal literature extensively discusses the applicability of current copyright law, and especially the fair use defence, to the use of protected content in AI training, and possible licensing and author remuneration mechanisms, there is limited debate on an equally important issue: how to ensure transparency, or sufficient information, around AI training data. Without knowing whether and how their content is being used in the AI training process, right holders are not able to try to enforce or license their rights in AI context. While there have been a number of policy proposals to improve data transparency in the US and other jurisdictions (e.g., the European Union (EU), the United Kingdom (UK)), there is limited understanding of their suitability.

This article starts by explaining the meaning of and need for AI data transparency in the copyright context and beyond. It then demonstrates that current and recently proposed AI data transparency measures, such as the public disclosure of datasets, publication of data summaries, and data certification, face challenges related to effectiveness, technical feasibility, legal costs and risks, and the protection of trade secrets and privacy. It then develops a new framework for AI data transparency underpinned by flexibility, accountability, and a holistic approach to

* Associate Professor in Law at Macquarie Law School, Macquarie University (Australia); Associate Senior Researcher at Lithuanian Centre for Social Sciences (Lithuania). This paper is an output of a project funded by the Australian Research Council DP250102392.

copyright law. This proposal is intended to ensure effective AI data transparency and that the copyright law framework is well-balanced and takes into consideration the interests of all stakeholders involved.

CONTENTS

- I. Introduction..... 524
- II. AI Transparency: Origin, Meaning, and Rationales 531
 - A. Transparency in the AI context..... 531
 - B. AI Data Transparency 534
 - C. Rationales for AI data transparency..... 535
 - 1. Copyright Law Reasons..... 536
 - 2. Other Reasons for Data Transparency 542
- III. Current Approaches to AI Data Transparency and the Remaining Challenges 546
 - A. Publication of Datasets 547
 - 1. Effectiveness 547
 - 2. Feasibility and Cost 550
 - 3. Privacy 553
 - 4. Trade Secrets..... 554
 - B. Summaries of Training Data..... 559
 - 1. Recent Initiatives 559
 - 2. Trade Secrets and Privacy..... 564
 - 3. Feasibility and Costs..... 565
 - 4. Effectiveness 567
 - C. Certification of Datasets 569
 - 1. Trade Secrets and Privacy..... 570

2. Feasibility, Cost and Effectiveness	571
IV. A Proposed Approach to AI Data Transparency	575
A. The FAH Approach.....	576
1. Flexible	576
2. Accountable	577
3. Holistic.....	579
B. The FAH Approach and Current Challenges to AI Data Transparency	582
1. Effectiveness.....	582
2. Feasibility and Costs.....	584
3. Protection of Trade Secrets and Privacy.....	586
V. Conclusion	588

I. INTRODUCTION

On May 2, 2023, the Writers Guild of America started a 147-day strike, one significant issue being the exploitation of writers’ materials in Artificial Intelligence (“AI”) training.¹ The same year, artists, content platforms, and music and press publishing organizations started multiple actions against Stability AI, Meta, Open AI, Microsoft, Google, and many other AI producers for

¹ Dani Anguiano & Lois Beckett, *How Hollywood Writers Triumphed over AI – and Why It Matters*, THE GUARDIAN (Oct. 1, 2023), <https://www.theguardian.com/culture/2023/oct/01/hollywood-writers-strike-artificial-intelligence> [https://perma.cc/6RJQ-TKG6]; Jake Coyle, *In Hollywood Writers’ Battle against AI, Humans Win (for Now)*, AP NEWS (Sep. 27, 2023), <https://apnews.com/article/hollywood-ai-strike-wga-artificial-intelligence-39ab72582c3a15f77510c9c30a45ffc8> [https://perma.cc/KUN4-6XX9].

unauthorized use of their content in AI training.² Currently, there more than forty lawsuits pending against AI producers³ claiming violations of copyright and other laws in the US⁴ and other jurisdictions,⁵ with first two decisions handed down in 2025.⁶ This upheaval in creative industries has invigorated academic and policy discussions on how the use of creative content in AI development projects should be reconciled with copyright law.⁷ Current discussions mostly

² Kate Knibbs, *Every AI Copyright Lawsuit in the US, Visualized*, WIRED, (Dec. 19, 2024), <https://www.wired.com/story/ai-copyright-case-tracker/> [<https://perma.cc/53NV-VCSU>]. Note, that there were at least two copyright-related lawsuits against AI companies before 2023. See Thomson Reuters Enter. Centre GmbH v. Ross *Intel*.Intelligence Inc., 694 F. Supp. 3d 467 (D. Del. 2023) (non generative-AI); J. Doe 1, et al. v. Github, Inc., No. 22-cv-06823-JST, 2024 WL 235217 (N.D. Cal. Jan. 22, 2024) (initiated by computer experts, not by creative industry).

³ In this paper, AI producer is defined as ‘an organization or entity that designs, develops, tests and deploys products or services that use one or more AI systems.’ See ISO/IEC 22989:2022 INFORMATION TECHNOLOGY—ARTIFICIAL INTELLIGENCE—ARTIFICIAL INTELLIGENCE CONCEPTS AND TERMINOLOGY § 5.19.3.1 (International Organization for Standardization & International Electrotechnical Commission 2022) [hereinafter ISO/EIC].

⁴ For a list of all relevant lawsuits in the U.S., see Knibbs, *supra* note 2.

⁵ See, e.g., *Getty Images (US) Inc. v. Stability AI Ltd.*, [2025] EWHC (Ch) 38 (UK); Peter Dalton et al., *Navigating representative actions: takeaways from Getty Images v Stability AI*, HERBERT SMITH FREEHILLS (Jan. 31, 2025) <https://www.herbertsmithfreehills.com/notes/ip/2025-01/navigating-representative-actions-takeaways-from-getty-images-v-stability-ai> (on file with author) (UK case); Michael Wittlinger, *AI Law update from Germany: GEMA sues OpenAI before Munich Court*, LEXOLOGY (Nov. 14, 2024), <https://www.lexology.com/library/detail.aspx?g=1315945f-6233-4fea-9ef1-a84861cb27d2> [<https://perma.cc/R4BR-4L58>].

⁶ *Bartzy v. Anthropic PBC*, 787 F. Supp. 3d 1007 (N.D. Cal. 2025); *Kadrey v. Meta Platforms, Inc.*, 788 F. Supp. 3d 1026 (N.D. Cal. 2025).

⁷ See, e.g., U.S. COPYRIGHT OFF., COPYRIGHT AND ARTIFICIAL INTELLIGENCE, PART 1: DIGITAL REPLICAS (July 2024), <https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-1-Digital-Replicas-Report.pdf> [<https://perma.cc/M5AV-82A4>]; U.S. COPYRIGHT OFF., COPYRIGHT AND ARTIFICIAL INTELLIGENCE, PART

focus on how current copyright exceptions, such as fair use in the US⁸ and text and data mining (TDM) in the EU and UK,⁹ apply in the AI context, and possibilities for right holders to license their content for AI training purposes.¹⁰ However, the issue that has gained relatively little academic

2: COPYRIGHTABILITY (Jan. 2025), <https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-2-Copyrightability-Report.pdf> [<https://perma.cc/M8TZ-7KVR>]; U.S. COPYRIGHT OFF., COPYRIGHT AND ARTIFICIAL INTELLIGENCE, PART 3: GENERATIVE AI TRAINING (May 2025), <https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf> [<https://perma.cc/MD5Z-3MQZ>]; SECRETARY OF STATE FOR SCIENCE, INNOVATION AND TECHNOLOGY, COPYRIGHT AND ARTIFICIAL INTELLIGENCE (Dec. 2024) [hereinafter *UK Copyright and Artificial Intelligence Consultation*], <https://www.gov.uk/government/consultations/copyright-and-artificial-intelligence/copyright-and-artificial-intelligence> [<https://perma.cc/KFD7-L835>] (information site for Central Government) (UK); Enrico Bonadio & Luke McDonagh, *Artificial Intelligence as Producer and Consumer of Copyright Works: Evaluating the Consequences of Algorithmic Creativity*, INTELL. PROP. Q. 112 (2020); Jane Ginsburg & Luke Budiardjo, *Authors and Machines*, 34 BERKELEY TECH. L. J. 343 (2019).

⁸ See, e.g., Pamela Samuelson, *Generative AI Meets Copyright*, 381 SCIENCE 158 (2023); Matthew Sag, *Copyright Safety for Generative AI*, 61 HOUS. L. REV. 295 (2023); Matthew Sag, *Fairness and Fair Use in Generative AI*, 92 FORDHAM L. REV. 1887 (2023); Daniel Rodriguez Maffioli, *Copyright in Generative AI Training: Balancing Fair Use through Standardization and Transparency* (Aug. 21, 2023). Available at SSRN: <https://papers.ssrn.com/abstract=4579322>.

⁹ See, e.g., João Pedro Quintais, *Generative AI, Copyright and the AI Act*, 56 COMPUTER L. & SEC. REV. 106, 107 (2025); 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).

¹⁰ See Rita Matulionyte, *Generative AI and Copyright: Exception, Compensation or Both?*, 134 INTELL. PROP. F. 33 (2023); Celeste Shen, *Fair Use, Licensing, and Authors’ Rights in the Age of Generative AI*, 22 NW. J. TECH. & INTELL. PROP. 157 (2024); Martin Senftleben, *Generative AI and Author Remuneration*, 54 IIC 1535 (2023).

attention is the need for transparency around AI training data in the copyright law context.¹¹

While the need for transparency around AI *generally* has been highlighted for years in various policy, governance, and academic papers,¹² copyright holders started urgently demanding transparency around AI technologies, especially AI training data, once they decided to challenge the use of their works in AI training process in courts.¹³ In order to prove that AI producers infringed their copyright when they used copyright-protected content to train AI modules without authorization, copyright holders first need to prove that their content was indeed used for this purpose. Since

¹¹ For a limited discussion, see Shayne Longpre et al., *Data Authenticity, Consent, and Provenance for AI Are All Broken: What Will It Take to Fix Them?*, MIT (Mar. 27, 2024), <https://mit-genai.pubpub.org/pub/uk7op8zs/release/2> [<https://perma.cc/MWC9-ZQ8G>]; A. Shaji George, T. Baskar, & Digvijay Pandey, *Establishing Global AI Accountability: Training Data Transparency, Copyright, and Misinformation*, 2 PUIRP 75 (2024); Maffioli, *supra* note 8.

¹² See, e.g., *Principles and Approach*, MICROSOFT, <https://www.microsoft.com/en-us/ai/principles-and-approach> [<https://perma.cc/Y76K-LAQL>] (last visited Feb. 10, 2026); *Recommendation on the Ethics of Artificial Intelligence*, UNESCO (2022), <https://unesdoc.unesco.org/ark:/48223/pf0000381137> (on file with author); *Ethics guidelines for trustworthy AI*, EUROPEAN UNION (Apr. 8, 2019), <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai> [<https://perma.cc/XGX3-V4LH>]; Joel Walmsley, *Artificial Intelligence and the Value of Transparency*, 36 AI & SOC'Y 585 (2021); Stefan Larsson & Fredrik Heintz, *Transparency in Artificial Intelligence*, INTERNET POL'Y REV., May 5, 2020, at 1, <https://policyreview.info/node/1469>.

¹³ Agata Mrva-Montoya, *Meta allegedly used pirated books to train AI. Australian authors have objected, but US courts may decide if this is 'fair use.'* THE CONVERSATION (Mar. 31, 2025), <http://theconversation.com/meta-allegedly-used-pirated-books-to-train-ai-australian-authors-have-objected-but-us-courts-may-decide-if-this-is-fair-use-253105> [<https://perma.cc/RD29-S2L8>] (detailing how the Australian Society of Authors has called for “transparency regarding which copyrighted works have been used for AI training and the purposes of that training”).

most AI producers do not disclose the training datasets they use,¹⁴ without transparency around AI training data, proving copyright infringement claims is especially difficult. Even if right holders do not want to pursue legal actions but rather prefer negotiating licensing agreements with AI producers, they need to know whether and how their content is or will be used in AI training process.

Right holders' calls for transparency around training data has already resulted in several policy initiatives. For example, in the US, California's legislation on *Generative Artificial Intelligence: Training Data Transparency*, adopted in 2024, requires AI producers provide summaries of training data.¹⁵ Several other legislative proposals are pending at a federal level.¹⁶ In Europe, data transparency duties in copyright law context were set under the EU Artificial Intelligence Act (EU AI Act),¹⁷ and the UK has committed to improve data transparency, too.¹⁸ Canadian

¹⁴ Thomas Heldrup, *Report on AI model providers' training data transparency and enforcement of copyrights*, RIGHTS ALLIANCE (Sept. 5, 2024), <https://rettighedsalliancen.dk/wp-content/uploads/2024/09/Report-on-AI-model-providers-training-data-transparency-and-enforcement-of-copyrights.pdf> [<https://perma.cc/3NWT-RLVP>].

¹⁵ Cal. Assemb. B. 2013, § 3111(a), 2023–24 Leg., Reg. Sess. (Cal. 2024) [hereinafter California Act AB 2013] (introduced by Irwin). This bill may be accessed at https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202320240AB2013 [<https://perma.cc/P62A-FZMQ>].

¹⁶ AI Foundation Model Transparency Act of 2023, H.R. 6881, 118th Cong. (2023); Generative AI Copyright Disclosure Act of 2024, H.R. 7913, 118th Cong. (2024) [hereinafter AI Foundation Model Transparency Act].

¹⁷ Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024, Laying Down Harmonised Rules on Artificial Intelligence, 2024 O.J. (L 1689) art. 53(1)(d) [hereinafter EU AI Act]; see generally Madalina Busuioc, Deirdre Curtin, & Marco Almada, *Reclaiming Transparency: Contesting the Logics of Secrecy within the AI Act*, 2 EUR. L. OPEN 79 (2023) (discussing the EU AI Act).

¹⁸ *UK Copyright and Artificial Intelligence Consultation*, supra note 7, at C.4.

authorities have supported a voluntary code of conduct that calls on model producers to “[p]ublish a description of the types of training data” a model uses,¹⁹ while UN bodies have recommended international regulations on “data rights that enshrine transparency.”²⁰ Although AI producers still tend to keep their datasets secret,²¹ most likely to avoid scrutiny, there are also AI industry initiatives for better data documentation and provenance as well as rights clearance.²²

Despite policy and industry initiatives on AI data transparency in the copyright law context, legal literature is scarce on the goals and limitations of the existing and proposed data transparency rules²³ and how these rules could be designed both to meet the legitimate interests of right holders and to properly consider other competing interests of AI producers and the general public.

The aim of this article is to demonstrate the limitations of the current measures to increase transparency around AI training data and to propose a new model

¹⁹ *Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems*, GOVERNMENT OF CANADA (Sept. 2023) [hereinafter *Voluntary Code*], <https://ised-isde.canada.ca/site/ised/en/voluntary-code-conduct-responsible-development-and-management-advanced-generative-ai-systems> (on file with author).

²⁰ UNITED NATIONS EXECUTIVE OFFICE OF THE SECRETARY-GENERAL (EOSG), *OUR COMMON AGENDA POLICY BRIEF 5 A GLOBAL DIGITAL COMPACT — AN OPEN, FREE AND SECURE DIGITAL FUTURE FOR ALL*, at 16 (May 24, 2023), <https://www.un-ilibrary.org/content/papers/10.18356/27082245-28> [https://perma.cc/SE56-SPVM].

²¹ Heldrup, *supra* note 14.

²² See, e.g., *Data & Trusted AI Alliance*, THE DATA & TRUSTED AI ALLIANCE (2025), <https://dataandtrustalliance.org/> [https://perma.cc/4HK6-RAHR]; *Data Provenance*, ARDC AUSTRALIAN RESEARCH DATA COMMONS (2024), <https://ardc.edu.au/resource/data-provenance/> [https://perma.cc/9B6B-59LM].

²³ See, e.g., George et al., *supra* note 11; Longpre et al., *supra* note 11; Maffioli, *supra* note 8.

framework to increase such transparency, which would balance the interests of all stakeholders involved. The article first reviews the origin and various meanings of transparency in the AI context, delineates it from related concepts such as AI explainability, and demonstrates where AI *data* transparency stands in broader academic and policy debates. Second, the article demonstrates the rationale for transparency around AI training data in the copyright context and beyond. Third, the article critically examines three current approaches to ensuring transparency around AI training data in the copyright law context—the publication of datasets, the provision of summary descriptions of datasets, and the certification of datasets. The analysis demonstrates that these solutions do not properly balance a variety of interests involved, such as effectiveness, privacy, feasibility, and trade secrets, and while they may be suitable for some AI datasets, they are not effective in relation to other ones. Finally, the article proposes a new approach for AI data transparency in the copyright law context, underpinned by flexibility, accountability, and a holistic approach to copyright law (FAH framework). It is argued that this approach is more likely to achieve the intended goals, protect privacy and trade secrets, and be feasible and cost-effective, and thus more appropriately balance the interests of copyright holders with those of AI producers and the general public.

This paper applies an international comparative perspective: it examines the topic by referring to copyright laws in the US, UK, and EU. Despite the differences in national copyright laws, it aims to develop a model for AI data transparency that could be adopted by different jurisdictions, with certain modifications as necessary. Finding a model framework that could apply across different jurisdictions is especially relevant, keeping in mind the global nature of AI industries. Harmonized rules around AI data transparency would make it easier for the international

AI industry to comply with them and ensure their effectiveness.

II. AI TRANSPARENCY: ORIGIN, MEANING, AND RATIONALES

Transparency is a long-standing legal concept, both in the government and private sectors. Transparency generally means being open and accountable, ensuring that information is readily available to the public and enabling informed participation and scrutiny of government or private sector actions.²⁴ Government transparency has been promoted through various laws, including freedom of information laws adopted by governments across the globe.²⁵ Transparency is also important in corporate governance generally, with specific transparency obligations imposed on, for instance, the financial sector and digital platforms.²⁶ It has gained renewed attention during the last decade with the emergence of AI technologies.

A. *Transparency in the AI context*

With the emergence of AI and algorithmic decision-making more broadly, AI transparency and explainability principles—which were initially used as synonyms—were introduced to respond to opaque, or black-box, AI systems (i.e., systems, the internal functioning and production of outputs of which, could not be understood either by laymen

²⁴ See Carolyn Ball, *What Is Transparency?*, 11 PUB. INTEGRITY 293 (2009); Greg Michener & Katherine Bersch, *Identifying Transparency*, 18 INFO. POLITY 233 (2013).

²⁵ See, e.g., The Freedom of Information Act, 5 U.S.C. § 552.

²⁶ See, e.g., Benjamin Fung, *The Demand and Need for Transparency and Disclosure in Corporate Governance*, 2 UNIVERSAL J. MGMT. 72 (2014); Leilasadat Mirghaderi, Monika Sziron, & Elisabeth Hildt, *Ethics and Transparency Issues in Digital Platforms: An Overview*, 4 AI 831 (2023).

or even by AI experts themselves).²⁷ Due to the technically opaque nature of many AI algorithms, such as Artificial Neural Networks (ANNs), stakeholders started demanding that the outcomes or decisions made by algorithms should be explainable and interpretable by stakeholders, such as AI users, especially in high stakes scenarios where algorithms may affect legal rights or health or have similarly significant effects.²⁸ Computer scientists have started developing various explainable AI (XAI) techniques to help explain AI decisions and outcomes.²⁹ While some commentators still sometimes refer to this as ‘transparency,’ the increasing consensus is that the explanation of AI outputs fall under the concept of ‘AI explainability.’³⁰

²⁷ For a general debate on AI opacity and its meaning see Charlotte A. Tschider, *Legal Opacity: Artificial Intelligence’s Sticky Wicket*, 106 IOWA L. REV. ONLINE 126 (2021); Joshua Krook et al., *A Systematic Literature Review of Artificial Intelligence (AI) Transparency Laws in the European Union (EU) and United Kingdom (UK): A Socio-Legal Approach to AI Transparency Governance*, 5 AI & ETHICS, 4069 (2025) (available at: <https://doi.org/10.1007/s43681-025-00674-z>); Warren J. von Eschenbach, *Transparency and the Black Box Problem: Why We Do Not Trust AI*, 34 PHILOS. & TECHNOL. 1607 (2021); Jenna Burrell, *How the Machine ‘Thinks’: Understanding Opacity in Machine Learning Algorithms*, BIG DATA & SOC’Y, Jun. 2016.

²⁸ See Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016, on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation), 2016 O.J. (L 119), at 1–88 arts. 13–15, 22 (requiring the provision of meaningful information about automated decision-making).

²⁹ See, e.g., Rudresh Dwivedi et al., *Explainable AI (XAI): Core Ideas, Techniques, and Solutions*, 55 ACM COMPUT. SURV. 194:1 (2023); Ambreen Hanif, Xuyun Zhang & Steven Wood, *A Survey on Explainable Artificial Intelligence Techniques and Challenges*, IEEE, Oct. 2021, at 81.

³⁰ ISO/IEC, *supra* note 3, § 3.5.7 (defining explainability as “property of an AI system to express important factors influencing the AI system results in a way that humans can understand”).

In parallel to demands for explainability of AI outcomes, calls emerged for more general transparency about AI systems, their development, and use. For instance, stakeholders started demanding information on how AI systems have been developed and function, including their training datasets, to be able to examine the safety of AI systems and to trust them.³¹ Others requested disclosure when AI is in use and/or information about how specific AI systems are being used,³² or whether an AI system has been used to develop specific content (e.g., whether video or text has been synthetically generated).³³ Increasingly, the provision of such information about AI systems was recognized to fall under a concept of ‘transparency,’ which is defined here as “property of a system that appropriate information about the system is made available to relevant stakeholders.”³⁴ AI transparency, along with AI explainability, have become key ethical AI principles entrenched in multiple international, regional, and national

³¹ E.g., *Transparency for Machine Learning-Enabled Medical Devices: Guiding Principles*, FDA (Jun. 13, 2024), <https://www.fda.gov/medical-devices/software-medical-device-samd/transparency-machine-learning-enabled-medical-devices-guiding-principles> [<https://perma.cc/UUG5-4PZN>].

³² See, e.g., Rita Matulionyte, *Increasing Transparency around Facial Recognition Technologies in Law Enforcement: Towards a Model Framework*, 33 INFO. & COMM’NS TECH. L. 66 (2024) (calling for more transparency around the use of facial recognition technologies by government).

³³ Dhruva Krishna, *Deepfakes, Online Platforms, and a Novel Proposal for Transparency, Collaboration, and Education*, 27 RICH. J.L. & TECH. 1, 51–63 (2020) (proposing transparency disclosures); Bart van der Sloot & Yvette Wagensveld, *Deepfakes: Regulatory Challenges for the Synthetic Society*, 46 COMPUT. L. & SEC. REV. 1, 7 (2022) (discuss whether transparency obligations under the proposed AI Act are sufficient).

³⁴ ISO/IEC, *supra* note 3, § 3.5.15.

documents,³⁵ and discussions are ongoing as to how they should be implemented with relation to different AI technologies applied in different sectors and contexts.

B. AI Data Transparency

While transparency around AI might include provision of information about various aspects of AI (e.g., its development, use, limitations), one part of the AI transparency principle is transparency around AI training data (or ‘AI data transparency’). Academic literature, policy documents, and industry standards are increasingly requiring AI producers to ensure transparency—or provide information—about AI training data.³⁶ Information about what training datasets were used to develop specific AI models and how training was performed will arguably help address various ethical concerns, as discussed below.³⁷

³⁵ See, e.g., *Recommendation of the Council on Artificial Intelligence*, OCEC (Mar. 5, 2024), <https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449> [<https://perma.cc/VU8S-PGAR>]; UNESCO, *supra* note 12; *Ethics and Governance of Artificial Intelligence for Health*, WORLD HEALTH ORGANIZATION [WHO] 26–27 (2021), <https://iris.who.int/server/api/core/bitstreams/f780d926-4ae3-42ce-a6d6-c898a5562621/content> (on file with author).

³⁶ E.g., California Act AB 2013, *supra* note 15; EU AI Act, *supra* note 17; DATA & TRUSTED AI ALLIANCE, *supra* note 22; ARDC, *supra* note 22.

³⁷ MOHAMMAD HAMDY & MONA HAMDY, THE UNSEEN LAYERS OF AI: AN EXPLORATION OF POOR DATA PROVENANCE IN MODEL TRAINING 2–5 (G20 Brasil, 2024) <https://papers.ssrn.com/abstract=49860> (on file with author); Christopher Yoo, *Beyond Algorithmic Disclosure For AI*, 25 COLUM. SCI. & TECH. L. REV. ONLINE 314, 318–321 (2024); Karl Manheim & Lyric Kaplan, *Artificial Intelligence: Risks to Privacy and Democracy*, 21 YALE J.L. & TECH. 106, 182 (2019) (“[R]equire full transparency on the downstream uses of user data”); Christina Han, *Parasites or Boosters of Human Creativity?: A Model to Solve Copyright Issues of Training Datasets of Artificial Intelligence before AlphaArt Defeats Human Artists*, 49 AIPLA Q. J. 281, 314 (2021) (advocating for increased data transparency of AI systems).

While legal and policy experts refer to ‘data transparency’ in the computer science domain this is associated with ‘data provenance’ initiatives.³⁸ The latter term is, however, broader than transparency around AI datasets and includes not only solutions designed to track the origin of training data, its owners, and/or licensing terms,³⁹ but also discussions on how (i.e., by which technical means) copyright holders could consent to the use of their content in AI training—or refuse such consent—as well as content-authenticity initiatives (e.g., technical measures allowing the labelling of AI-generated outputs, including AI-generated deepfakes).⁴⁰ This paper focuses only on certain data provenance initiatives, namely, those that aim to increase information about AI training data, its sources, owners, and other information relevant from the copyright law perspective.

C. Rationales for AI data transparency

The rationales for AI transparency *generally* have been well covered in literature.⁴¹ Different stakeholders—AI users, impacted persons, AI deployers, regulators, legal experts, policymakers, and others—need different types of information about AI systems for different purposes, such as for ensuring their quality, safety, legality, and alignment with societal and ethical norms, increasing trust and safe use of these technologies, and enabling the ability to contest AI

³⁸ Longpre et al., *supra* note 11 at 12 (discussing data provenance as a way to ensure AI transparency); Hamdy, *supra* note 37 at 2–5.

³⁹ E.g., Shayne Longpre et al., *The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI* (2023) (unpublished manuscript), <https://hal.science/hal-04290233> [<https://perma.cc/JK9W-P5ZD>] (HAL Id: hal-04290233v1).

⁴⁰ See Longpre et al., *supra* note 11.

⁴¹ See, e.g., Heike Felzmann et al., *Towards Transparency by Design for Artificial Intelligence*, 26 SCI ENG ETHICS 3333 (2020).

outputs.⁴² AI *data* transparency is required for diverse reasons too, which could be classified into copyright law reasons and other reasons.⁴³

1. Copyright Law Reasons

Transparency around AI training data is crucial for copyright holders to be able to exercise and enforce their rights. Many AI models, including generative AI models, have arguably been trained on large amounts of content protected by copyright,⁴⁴ and this has often (or almost always) been done without the authorization from right holders.⁴⁵ While it is yet to be fully affirmed by courts, unauthorized use of content in AI training process would amount to copyright infringement, at least in some jurisdictions, as exceptions such as fair use in the US or

⁴² For a full taxonomy of AI transparency stakeholders and their needs, see ISO/EIC, *supra* note 3, § 5.19.3.1.

⁴³ See, e.g., Tschider, *supra* note 27, at 700–704 (discussing transparency for AI safety, fairness and responsibility); Krook, *supra* note 27, at 10–14; Quintais, *supra* note 9, at 10 (demonstrating that transparency is important “to privacy and data protection, research, the prohibition of discrimination and respect for diversity (related e.g., to the avoidance or mitigation of bias), and fair competition”).

⁴⁴ E.g., Guangkai Xu et al., *Diffusion Models Trained with Large Data Are Transferable Visual Models*, ARXIV (Mar. 15, 2024), <https://arxiv.org/html/2403.06090v2> [<https://perma.cc/Z87Q-EZHA>] (“Stable Diffusion is pretrained on the massive LAION-5B dataset, around 5 billion text-image pairs.”).

⁴⁵ In some cases, companies secure licenses to the content created by their users. For example, Meta’s Terms of Service allow Meta to use Facebook users’ content in any way. See *Meta Terms of Service*, META (Jan. 1, 2025) <https://www.facebook.com/terms/> [<https://perma.cc/TZQ2-Y9WG>] (“[Y]ou grant us a non-exclusive, transferable, sub-licensable, royalty-free, and worldwide license to host, use, distribute, modify, run, copy, publicly perform or display, translate and create derivative works of your content[.]”). In many other cases, AI producers scrape content from online sources without authorization from right holders.

TDM exceptions in the EU and UK are unlikely to apply in all cases.

For instance, while in two recent US judgements the use of content in the training of generative AI was found to be fair use in certain circumstances,⁴⁶ judges in both cases indicated that fair use will not apply in certain other AI training scenarios.⁴⁷ In the EU, while there are two TDM exceptions that will apply for the use of content in an AI training context,⁴⁸ copyright infringement will be found if right holders opt out from the use of their works in TDM (or commercial AI training projects),⁴⁹ but the AI producer nevertheless uses their content disregarding this opt out.⁵⁰ The UK has a limited TDM exception for non-commercial use, which does not apply to commercial TDM scenarios.⁵¹ Thus, the use of content in commercial AI projects in the UK is likely to amount to copyright infringement.⁵² Finally, many countries around the world do not have exceptions that

⁴⁶ See *Bartz v. Anthropic PBC*, 787 F. Supp. 3d 1007, 1034 (N.D. Cal. 2025); *Kadrey v. Meta Platforms, Inc.*, 788 F. Supp. 3d 1026, 1059–60 (N.D. Cal. 2025).

⁴⁷ See *Bartz*, 787 F. Supp. 3d at 1034 (finding that fair use will not apply when training content is downloaded from pirate websites for multiple uses); *Kadrey*, 788 F. Supp. 3d at 1060 (suggesting that fair use would not have been established if plaintiffs demonstrated dilution).

⁴⁸ 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, 2019 O.J. (L 130), 92–125, arts. 3–4 [hereinafter EU DSM Directive].

⁴⁹ See EU AI Act, *supra* note 17, recitals 105, 107 (clarifying that TDM is considered to be a part of the AI training process, and implicitly confirming that TDM exceptions apply to AI training).

⁵⁰ See EU DSM Directive, *supra* note 48, art. 4(3).

⁵¹ Copyright, Designs and Patents Act 1998, c. 48, § 29A (UK).

⁵² See *Getty Images (US) Inc. v. Stability AI Ltd.*, [2025] EWHC (Ch) 38 (UK); see generally Matthew Coulter, *Aiming for Fairness: an Exploration into Getty Images v. Stability AI and its Importance in the Landscape of Modern Copyright Law*, 34 DEPAUL J. ART TECH. & INTELL. PROP. L 124 (2024) (discussing AI in the context of copyright law).

are likely to apply in AI data training scenarios, in which case the establishment of copyright infringement is even more likely.⁵³

As identified above, one of the most significant challenges for right holders seeking to enforce their rights in AI training scenarios is a lack of transparency around datasets used to train AI models. Some AI companies, such as Stability AI, have disclosed the datasets they used to train some of their models,⁵⁴ however, training datasets of many other AI models, especially more recent ones, are often not known.⁵⁵ AI producers, such as IBM, Google, and Facebook, claim that training datasets constitute their proprietary information and are thus protected as trade secrets,⁵⁶ which creates one of the greatest challenges to AI data

⁵³ See, e.g., Rita Matulionyte, *Australian Copyright Law Impedes the Development of Artificial Intelligence: What Are the Options?*, 52 IIC 417 (2021) (explaining that Australia lacks a fair use or text-and-data-mining exception).

⁵⁴ See, e.g., Heldrup, *supra* note 14, at 45–48 (discussing how Stability AI has disclosed that first versions of Stable Diffusion have been trained on LAION-2B and LAION-5B datasets).

⁵⁵ See, e.g., Heldrup, *supra* note 14, at 23 (noting that while Open AI disclosed some information about GPT2 and GPT-3 training data, they did not disclose any information about GPT-4 training data); see also Han, *supra* note 37, at 305–08; Shlomit Yanisky-Ravid & Sean K. Hallisey, “Equality and Privacy by Design”: A New Model of Artificial Intelligence Data Transparency via Auditing, Certification, and Safe Harbor Regimes, 46 *FORDHAM URB. L.J.* 438, 437 (identifying the lack of transparency on “the data sources” of an AI system as one of “problematic features of machine learning algorithms that make regulation difficult”).

⁵⁶ Han, *supra* note 37, at 305; see also INTERNATIONAL TECHNOLOGY LAW ASSOCIATION, *RESPONSIBLE AI: A GLOBAL POLICY FRAMEWORK* 277–78 (2019), https://www.itechlaw.org/sites/default/files/Responsible_AI.pdf [<https://perma.cc/4L8X-VJTL>] (noting the tension between protecting the effort and resources that go into building AI, and the contrary goal of maintaining fairness and transparency in AI, especially in cases of determining liability).

transparency.⁵⁷ As a result, copyright holders have to rely on indirect evidence to prove that their works have indeed been used in AI training process, such as previous versions of the trained AI model or via outputs of the AI model that reproduce substantial parts of the pre-existing works.⁵⁸

During litigation, AI training data could potentially be found during the discovery process, but discovery proceedings are expensive and unaffordable for many right holders, particularly when trade secrets are involved, which can further complicate the process.⁵⁹ In addition, there has been no legal duty to properly document training data or maintain it. Thus, at the time of litigation, AI companies might have no appropriately-recorded information about all the training data they used or might have destroyed their datasets, to avoid liability or for other reasons.⁶⁰ When

⁵⁷ Han, *supra* note 37, at 306; *see also* Quinn Emanuel Urquhart & Sullivan, LLP, *April 2020: The Increasing Importance of Trade Secret Protection for Artificial Intelligence*, JD SUPRA (Apr. 27, 2020), <https://www.jdsupra.com/legalnews/april-2020-the-increasing-importance-of-64465/> [<https://perma.cc/JHZ5-YTUK>] (pointing out that as the importance of AI continues to grow, companies increasingly seek to protect their AI assets as trade secrets).

⁵⁸ *See, e.g.*, Complaint at 85–93, *The New York Times Co. v. Microsoft Corp.*, 777 F. Supp. 3d 283 No. 1:23-cv-11195 (S.D.N.Y. 2025) (noting that New York Times, in their complaint against OpenAI, refer to training datasets of early GPT models (GPT-2 and GPT-3) as no information about GPT-4 is available).

⁵⁹ *See* Maverick Law Firm Team, *Protection of Trade Secrets from Litigation Discovery*, MAVERICK LAW FIRM (Feb. 16, 2018), <https://www.mavericklaw.com/blog/protection-trade-secrets-litigation-discovery/> [<https://perma.cc/9X6D-QTMC>] (citing two decisions which demonstrate the difficulty in conducting discovery in trade secret disputes).

⁶⁰ *See* Martin Kretschmer, Thomas Margoni, & Pinar Oruç, *Copyright Law and the Lifecycle of Machine Learning Models* 55 IIC 110, 125 (2024) (“[C]ommercial AI developers are told by their legal departments to ‘mine everything and then destroy the training material’ since it will be very difficult to reverse-engineer the trained model, go back to the training material and prove infringement.”).

datasets are not available anymore, it is generally impossible to know what data was used for training purposes, (i.e., it generally cannot be identified by examining the trained AI model or by trying to reverse engineer it).⁶¹ Further, discovery of training data during litigation is not a suitable option for right holders who do not intend to litigate but rather want to start licensing negotiations. Overall, transparency around training data is essential for copyright holders to be able to exercise their rights (i.e., to license them to AI producers or enforce them when their content is used for training purposes without authorization).⁶²

Some commentators argue that legal requirements for AI data transparency in the copyright law context cannot be conceptually justified as the use of content for AI training does not infringe copyright.⁶³ Conceptually, AI producers could be required to provide certain information about the training data only if there is a legitimate interest of another party to access such information. These commentators argue that the use of works in AI training is covered by the fair use exception, and therefore no permission or fee is required for the use of such works.⁶⁴ Thus, there is no conceptual justification—no legitimate interest—for right holders to demand information about the use of their works in training data.⁶⁵

⁶¹ Han, *supra* note 37, at 307 n.142 (“The collected data are often converted to a different format and are optimized for machine learning, thus often making it difficult to reconstruct the original form of the collected data from the processed data.”).

⁶² Heldrup, *supra* note 14, at 4.

⁶³ MATTHEW SAG, PAMELA SAMUELSON, & CHRISTOPHER JON SPRIGMAN, COMMENTS IN RESPONSE TO THE COPYRIGHT OFFICE’S NOTICE OF INQUIRY ON ARTIFICIAL INTELLIGENCE AND COPYRIGHT 33–35 (2024), <https://papers.ssrn.com/abstract=4976391> (on file with author).

⁶⁴ *Id.*

⁶⁵ *Id.*

However, as has been argued above, unauthorized use of copyright-protected content in AI training is likely to constitute copyright infringement in many jurisdictions, at least in some cases.⁶⁶ At the same time, the copyright landscape is in flux: courts in different countries are examining the limits of currently existing exclusive rights and exceptions as they apply in the context of AI, while lawmakers are considering various copyright law reform options.⁶⁷ If lawmakers or courts determine that the use of protected subject matter during AI training does not require authorization from right holders, then indeed calls for data transparency would lose their conceptual basis. However, the most recent court decisions⁶⁸ and policy proposals to remunerate right holders, at least in some situations,⁶⁹ seem to indicate the willingness of governments to allow right holders to maintain certain control over—or remuneration for—the use of their works in the AI development process. If this is the case, copyright holders will retain their

⁶⁶ See *supra* Section II.C.

⁶⁷ See Knibbs, *supra* note 2 (listing cases pending against AI producers); see, e.g., *Artificial Intelligence and Copyright*, 88 Fed. Reg. 59942, 59946 (Aug. 30, 2023) (disclosing government consultations on AI and copyright issues and questions on AI training and copyright).

⁶⁸ See *Bartz v. Anthropic PBC*, 787 F. Supp. 3d 1007 (N.D. Cal. 2025) (establishing copying of works for AI training purposes from pirate sites as a copyright infringement); *Kadrey v. Meta Platforms, Inc.*, 788 F. Supp. 3d 1026 (N.D. Cal. 2025) (finding infringement would have been established if plaintiffs provided evidence on dilution of market value); see also *Thomson Reuters Enter. Centre GmbH v. Ross Intelligence Inc.*, 694 F. Supp. 3d 467 (D. Del. 2023).

⁶⁹ Boriana Guimberteau, *AI, Machine Learning and Big Data Laws and Regulations 2024*, STEPHENSON HARWOOD (Jun. 17, 2024) <https://www.stephenonharwood.com/insights/ai-machine-learning-and-big-data-laws-and-regulations-2024?08042025034035> [https://perma.cc/U5JB-GCR8] (noting that French Parliament proposed a law that “aims at protecting authors, allowing them to perceive fair and equitable remuneration, and encouraging innovation and promoting artistic diversity”).

legitimate interest in transparency around AI training data, and the conceptual justification for such calls cannot be meaningfully challenged.

2. Other Reasons for Data Transparency

Apart from copyright-related reasons, there are numerous other reasons to strengthen transparency around AI training data, including bias, privacy, and liability considerations. Firstly, experts have called for transparency around AI training data to help address bias and discrimination risks that some algorithms have demonstrated.⁷⁰ Biased algorithms may raise safety issues, since biased AI decisions or outcomes might harm human health or violate legal rights or other legitimate interests. Biased AI decisions or outcomes might lead to unfair decisions and perpetuate and amplify bias and discrimination found in human decisions and our society more generally.⁷¹ Biased or discriminatory AI models and their outputs are, in many cases, caused by biased data,⁷² i.e., data that is not “broad enough to be representative of the expected operational data input.”⁷³ Therefore, transparency

⁷⁰ See Anne L. Washington, *How to Argue with an Algorithm: Lessons from the COMPAS-ProPublica Debate*, 17 COLO. TECH. L.J. 131, 154–55 (2018) (requiring provenance around predictive algorithms in judicial sector); see also Hamdy, *supra* note 37, at 2–5; Yoo, *supra* note 37, at 318; Peter K. Yu, *The Algorithmic Divide and Equality in the Age of Artificial Intelligence*, 72 FLA. L. REV. 331, 373 (2020); Karl Werder, Balasubramaniam Ramesh, & Rongen (Sophia) Zhang, *Establishing Data Provenance for Responsible Artificial Intelligence Systems*, ACM TRANS. MANAGE. INF. SYST., Mar. 10, 2022, at 22:1, 22:1.

⁷¹ See, e.g., Brandon L. Garrett & Cynthia Rudin, *Interpretable Algorithmic Forensics*, PNAS, OCT. 2, 2023, at 1.

⁷² ISO/IEC, *supra* note 3, § 5.15.9.

⁷³ *Id.* at § 5.10. Algorithms might be biased due to other reasons (e.g., algorithmic structure or parameters) and the impacts of algorithms might be discriminatory due to the bias of their users (social bias). See Marcus Smith & Monique Mann, *Facial Recognition Technology and Potential for Bias and Discrimination*, in THE CAMBRIDGE HANDBOOK OF FACIAL

around training data is essential for stakeholders (especially experts and regulators) to assess the possible risks of AI bias and ensure that they are eliminated or reduced.

Lack of transparency around AI training data also poses risks to the privacy of individuals.⁷⁴ Many AI models are developed on large amounts of data, often scraped from online sources, which often include personal data acquired without authorization or consent from data subjects.⁷⁵ Without AI data transparency, individuals are not able to know whether their data has been collected, how it was used, who holds it, and provide or deny consent for such uses, while privacy authorities are not able to investigate privacy breaches. A notorious example of privacy violations in AI context is the Clearview AI case where a facial recognition system was trained with billions of facial images scraped from social media and other sites without authorization from individuals.⁷⁶ Authorities in the US, Europe, and Australia have found that such use of personal data in the AI training process has breached privacy laws.⁷⁷ While in the Clearview

RECOGNITION IN THE MODERN STATE 87 (Rita Matulionyte & Monika Zalnieriute eds., 2024).

⁷⁴ Krook, *supra* note 27, at 12; Hamdy, *supra* note 37, at 2; Manheim, *supra* note 37; Bofeng Pan, Natalia Stakhanova & Suprio Ray, *Data Provenance in Security and Privacy*, 55 ACM COMPUT. SURV. 323:1 (2023).

⁷⁵ E.g., JENNIFER KING & CAROLINE MEINHARDT, *RETHINKING PRIVACY IN THE AI ERA: POLICY PROVOCATIONS FOR A DATA-CENTRIC WORLD*, STANFORD UNIVERSITY (2024), <https://hai.stanford.edu/policy/white-paper-rethinking-privacy-ai-era-policy-provocations-data-centric-world> [<https://perma.cc/8AVR-HB6W>].

⁷⁶ For a general discussion of the Clearview case, see Miriam Kohn, *Clearview AI, TikTok, and the Collection of Facial Images in International Law*, 23 CHI. J. INT'L L. 195 (2022).

⁷⁷ See Kashmir Hill, *Twitter Tells Facial Recognition Trailblazer to Stop Using Site's Photos*, THE NEW YORK TIMES (Jan. 22, 2020), <https://www.nytimes.com/2020/01/22/technology/clearview-ai-twitter-letter.html> [<https://perma.cc/CYT7-H6C5>] (discussing how Clearview

AI case it was not disputed that personal data (facial images) were used to train the system, in other cases, the use of personal data might be impossible to ascertain unless AI producers are transparent about the training data they use.⁷⁸

Further, information about training data and the training process might be important when allocating liability for harm caused by AI systems. Commentators have argued that it is often very difficult to determine who is accountable for the harm caused by an AI tool—the AI producer, AI deployer, or AI user—especially if the AI decision, its rationale, and how it was generated, cannot be explained.⁷⁹ For example, when an autonomous vehicle killed a pedestrian crossing a street, the inability to explain the internal workings of AI made it very difficult to attribute liability to any of the parties involved.⁸⁰ Access to AI training data can be important in some cases in solving this liability conundrum. For example, if analysis of the AI

AI was trained on millions of photos scraped from social media sites in violation of their terms of use).

⁷⁸ GPT-4 was trained on a large amount of data, including data that scraped from online sources that may contain personal information; however, this cannot be confirmed until detailed information about training data is made transparent. *See e.g.*, Maximilian Schreiner, *GPT-4 architecture, datasets, costs and more leaked*, THE DECODER (Jul. 11, 2023), <https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/> [<https://perma.cc/K8X4-9EN7>] (suggesting that GPT-4’s “training data included CommonCrawl & RefinedWeb”).

⁷⁹ *See* Krook, *supra* note 27, at 11; *see also* Burrell, *supra* note 27; Robert Van Krieken, *The Organization of Ignorance: The Australian “Robodebt” Affair, Bureaucracy, Law and Politics*, 50 CRITICAL SOCIOLOGY 1379, 1381 (2024).

⁸⁰ *See* Daisuke Wakabayashi, *Self-Driving Uber Car Kills Pedestrian in Arizona, Where Robots Roam*, N.Y. TIMES (Mar. 19, 2018), <https://www.nytimes.com/2018/03/19/technology/uber-driverless-fatality.html> [<https://perma.cc/4SQZ-LWHN>]. Similar accountability challenges have been identified in other high-risk settings, such as healthcare. *See* Rita Matulionyte, Paul Nolan, Farah Magrabi, & Amin Beheshti, *Should AI-Enabled Medical Devices be Explainable?*, 30 INT’L J.L. & INFO. TECH. 151, 172–73 (2022).

training dataset shows that the dataset did not contain a sufficient amount and/or diversity of images of a person crossing a street, and this has caused the AI system to not correctly identify a pedestrian, this can be a strong piece of evidence suggesting that AI producer is at fault, (i.e., they have not ensured that AI was trained on an appropriate dataset.)⁸¹

Commentators also argue that AI transparency, including AI data transparency, is essential for the public to trust and adopt AI technology,⁸² especially in higher risks settings. For example, in the medical context, if clinicians and patients do not understand on what type of data a specific AI system was trained and whether the data from local patients' populations has been used, they will arguably not trust using these AI devices in their own clinical settings and will not adopt them in practice.⁸³

In addition, transparency around AI training data would be particularly important when data or algorithms are repurposed from one function to another, as training data can show where gaps exist that could result in unexpected outcomes.⁸⁴ Culturally, the lack of transparency around training data makes it difficult to assess the model representativeness, which could threaten linguistic and cultural diversity and perpetuate the exclusion of certain cultures or groups.⁸⁵

⁸¹ Cf. Matulionyte, *supra* note 80, at 172–73 (arguing that AI explainability has a limited or unclear role in allocating liability).

⁸² Krook, *supra* note 27, at 12.

⁸³ See Onur Asan, Arpaslan Bayrak & Avishek Choudhury, *Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians*, 22 J. MED. INTERNET RES. 4 (2020); Nicholas Diakopoulos, *Accountability in Algorithmic Decision Making*, 59 COMMUN. ACM 56, 61 (2016).

⁸⁴ Yoo, *supra* note 37, at 321.

⁸⁵ Hamdy, *supra* note 37, at 2–5.

III. CURRENT APPROACHES TO AI DATA TRANSPARENCY AND THE REMAINING CHALLENGES

As AI data transparency has been requested for diverse reasons, numerous different solutions have been adopted or proposed to ensure it.⁸⁶ In the context of copyright law, three main approaches to data transparency have been discussed: the disclosure of training datasets, the publication of data summaries, and the certification of data. While these approaches have certain strengths, they also have various challenges, and their ability to balance various involved interests is limited.

This section examines the suitability of the available approaches to AI data transparency based on four criteria. First, will the proposed measures likely be effective in achieving their goal, namely, will they result in the disclosure of information that is sufficient for right holders to exercise (enforce or license) their rights? Second, are current approaches technically feasible without putting an unreasonable financial burden on AI producers? Third, do they adequately protect trade secrets of AI producers? Finally, do they appropriately protect the privacy of individuals whose data was included in training datasets? While the first criterion (effectiveness) takes into account the right holders' needs, the second and the third (technical feasibility, costs, and trade secrets) consider AI producers' interests, and the final criterion (privacy) relates to public interest concerns. Together, they will allow for the identification of whether the analyzed measures appropriately address the needs and concerns of all involved stakeholders and properly balance all interests involved.

⁸⁶ See e.g., Longpre et al., *supra* note 11; Yanisky-Ravid, *supra* note 55.

A. *Publication of Datasets*

The first option to ensure transparency around training datasets, as demanded by copyright holders, is to require AI producers to make training datasets public by, for instance, publishing them online.

1. Effectiveness

Many training datasets, such as LIAON5B or ImageNet,⁸⁷ are publicly available and can be accessed and analyzed by any interested party, including any right holder. Organizations and individuals have been publishing large and small datasets on public repositories or dataset libraries such as Kaggle, UCI Machine Learning Repository, Google Dataset Search, and others.⁸⁸ Many AI producers have been using these publicly available datasets to develop their AI models and some have disclosed to the public which training datasets they used. For example, Stability AI has previously announced that Stable Diffusion has been trained on the LIAON5B dataset, which is a publicly available dataset.⁸⁹ EleutherAI, which developed GPT-NeoX-20B, provides direct access to a copy of its training data, along with general descriptions of its datasets.⁹⁰

When an AI producer discloses what datasets were used and makes them publicly available, right holders can

⁸⁷ For more about available datasets, see *List of datasets for machine-learning research*, WIKIPEDIA (Mar. 29, 2025), https://en.wikipedia.org/w/index.php?title=List_of_datasets_for_machine-learning_research&oldid=1282902839.

⁸⁸ For a list of the most popular dataset libraries and repositories, see *65 of the Best Training Datasets for Machine Learning*, SMARTONE.AI (Mar. 25, 2023), <https://smartone.ai/blog/65-of-the-best-training-datasets-for-machine-learning> [<https://perma.cc/T7ZP-ED6Q>].

⁸⁹ See Heldrup, *supra* note 14, at 46–49.

⁹⁰ *Id.* at 38. For a technical paper describing the dataset, see Leo Gao et al., *The Pile: An 800GB Dataset of Diverse Text for Language Modeling*, ARXIV (Dec. 31, 2020), <http://arxiv.org/abs/2101.00027> [<https://perma.cc/S79R-J945>].

search the datasets to find out, at least, whether their content was used as part of that dataset. However, while this option of providing transparency can be helpful,⁹¹ it has its limitations, too. First, searching through datasets will, in most cases, require certain technical expertise and resources. Each model might be trained on multiple datasets, including large datasets containing huge amounts of content scraped from online sources and curated datasets. Technical expertise would be required to decode, scrutinize, and comprehend such datasets.⁹² Thus, while searching published datasets might be an option for large right holders who can afford to hire experts to conduct such searches, this might be less feasible or impossible for small or individual right holders without expertise and/or resources. Datasets or content items in the dataset might have very limited metadata, which might make it difficult for right holders to find whether their content (and which item(s) in particular) is a part of a dataset.

Some dataset creators facilitate dataset searches by providing an interface for right holders to identify whether their work was included in the dataset. For instance, LIAON datasets can be searched on the ‘Have I been trained’ platform,⁹³ and Atlantic has recently published a search tool allowing authors to search for their books in the LibGen dataset.⁹⁴ Such interfaces are, unfortunately, not available for many datasets.

⁹¹ See Heldrup, *supra* note 14 (providing reports that examine the transparency of datasets and, where a dataset is published, identifying which right holder content has been used and whether from licensed sources).

⁹² Maffioli, *supra* note 8, at 41; Heldrup, *supra* note 14, at 32.

⁹³ See Have I been Trained?, SPAWNING AI, <https://havebeentrained.com/> [<https://perma.cc/7BTG-LBEY>] (last visited Feb. 14, 2026).

⁹⁴ See Alex Reisner, *Search LibGen, the Pirated-Books Database That Meta Used to Train AI*, THE ATLANTIC (Mar 20, 2025),

Another problem is that, even if a right holder manages to identify their work in the dataset, there might be challenges to verify the licensing status of the work. The description of the dataset or metadata attached to a specific content piece will rarely (if ever) provide information about the copyright, ownership, or licensing status. This information is important both for right holders wishing to understand whether their content was taken from a properly licensed source and AI producers wishing to ensure that the rights to content they use in AI training have been properly cleared.

For instance, Common Crawl—a text dataset which contains data scraped from internet—is very commonly used to pretrain text-based AI models.⁹⁵ For each content item, it provides the website address (e.g., URL) from which the content is taken, scrape times, and the raw documents taken from online sources.⁹⁶ However, this metadata is unlikely to identify the content’s author, the title of the work or any information about copyright and licensing status.

Some data libraries provide more detailed information about data provenance. For instance, Hugging Face Datasets, a very popular data library for AI training, has integrated ‘data cards’ and information on whether right holders have provided consent to use the content.⁹⁷ These

<https://www.theatlantic.com/technology/archive/2025/03/search-libgen-data-set/682094/> [<https://perma.cc/S5CH-E5ZT>].

⁹⁵ See COMMON CRAWL, <https://commoncrawl.org/> [<https://perma.cc/A7LN-TGG6>]. 60 percent of GPT-3 training data is from Common Crawl. See Liz O’Sullivan & John P. Dickerson, *Here are a few ways GPT-3 can go wrong*, TECHCRUNCH (Aug. 7, 2020), <https://techcrunch.com/2020/08/07/here-are-a-few-ways-gpt-3-can-go-wrong/> [<https://perma.cc/CG9N-Q95T>].

⁹⁶ It also follows the robots.txt exclusion protocol, and arguably even adheres to requests to remove content from the dataset. See Longpre et al., *supra* note 11.

⁹⁷ Collecting consent and opt-out requests is enabled by, for example, Spawning. See *Spawning.Ai Announces to Have Collected Opt-out*

measures encourage documentation of data provenance and enable users to filter for content with consent from right holders, which is useful for AI producers wishing to use only licensed data and for right holders to check whether datasets contain only properly licensed content. However, recent research shows that despite these measures, the documentation of data provenance in most datasets is uneven and often incorrect, as this information is loosely crowdsourced.⁹⁸ Some projects, such as the Data Provenance Initiative, have tried to add more comprehensive and accurate structured information to the most popular textual datasets (e.g., lineage of sources, creators, licenses), however, this requires significant human labor and comes at high costs, and thus, is difficult to scale.⁹⁹ This leads us to the feasibility and costs issue.

2. Feasibility and Cost

In terms of technical feasibility and cost, publishing the dataset online (e.g., on a company's website or public repository) is generally feasible and does not involve significant costs as public repositories charge fees that are affordable even to small companies.¹⁰⁰ In addition, if an AI producer uses publicly available datasets, disclosing them to the public would require only mentioning the name of the dataset and its online location, and the interested

Requests for 80 Million Artworks, OPEN FUTURE (Mar. 8, 2023), <https://openfuture.eu/note/spawning-ai-announces-to-have-collected-opt-out-requests-for-80-million-artworks> [<https://perma.cc/P8NJ-P9QG>].

⁹⁸ Longpre et al., *supra* note 11, at 13.

⁹⁹ Longpre et al., *supra* note 11, at 13.

¹⁰⁰ Public repositories might include a fee per dataset for an online repository to cover its operational costs and storage costs, which are normally not high and usually depend on the size of the dataset. *See, e.g., Pricing*, HUGGING FACE, <https://huggingface.co/pricing> [<https://perma.cc/4J9G-CAU2>] (last visited Feb. 14, 2026).

stakeholders could access and/or download it from the public repository, with no costs to the AI producer.

However, ensuring that a dataset contains all metadata required by right holders is a very complex task, the feasibility of which could be questioned.¹⁰¹ Indeed, if an AI producer is required not to merely disclose the datasets used, but also provide information demonstrating that copyright in datasets—and possibly each item in the dataset—have been cleared, this might not be technically (or economically) feasible. AI producers might use multiple datasets that are packaged and repackaged into large collections. Data might not be annotated, or annotations might be lost during this process, so it might be impossible to know to whom each item belongs and under what licensing terms they were accessed.¹⁰² Also, it is questionable whether AI producers always properly document what they use for training purposes as data documentation and provenance standards are only developing, and binding requirements to document data are only emerging.¹⁰³ Further, after the training process is completed, they might not keep the training data but delete it instead, so they might not have the data when it is demanded by right holders.¹⁰⁴ Additionally, in some cases AI producers never get hold of the data. For instance, when

¹⁰¹ Exact costs will depend on factors such as the amount and nature of the data, the complexity of the annotation process, and the composition and location of the ML team. *See, e.g.,* Gurpreet Singh, *Machine Learning Development Cost: A Comprehensive Guide*, DEBUT INFOTECH (Jan. 17, 2025), <https://www.debutinfotech.com/blog/machine-learning-development-cost-analysis> [<https://perma.cc/EL4X-38WX>].

¹⁰² *See* Longpre et al., *supra* note 11, at 13.

¹⁰³ One of the first legislative instruments establishing training data documentation duties is the EU AI Act. *See* EU AI Act, *supra* note 17, art. 53 (requiring that developers of general-purpose AI systems provide documentation about the training and testing processes and the evaluation of results).

¹⁰⁴ *See* Kretschmer, *supra* note 60, at 125.

AI producers use a federated learning approach,¹⁰⁵ they might train their algorithms on data owned and held by third parties, and only receive the trained algorithm, but never data itself.

Tracking the copyright and licensing status of all data is especially complicated, if it is at all possible. As mentioned above, currently, publicly available datasets do not normally have information about the copyright and/or licensing status of individual content items in the dataset. The datasets will most frequently contain a short description of the dataset and a website from which that specific item was scraped without providing information relevant to determining copyright or licensing status.¹⁰⁶ Even those few datasets that claim to hold licensed data have been demonstrated to provide licensing information that is often not complete or accurate.¹⁰⁷ While data provenance initiatives are emerging to provide more information about each content item in selected datasets, this requires human labor and is very costly.¹⁰⁸

Even if techniques to provide data transparency improve, they will be associated with higher costs resulting from labelling each content item with information required by right holders (copyright and licensing status), and verifying that the metadata is correct.¹⁰⁹ These costs might increase the overall AI development costs.¹¹⁰ In addition,

¹⁰⁵ For more about federated learning, see Priyanka Mary Mammen, *Federated Learning: Opportunities and Challenges*, ARXIV (Jan. 14, 2021), <http://arxiv.org/abs/2101.05428> [<https://perma.cc/U9GN-K94X>] (explaining the working of federated learning).

¹⁰⁶ For a general discussion on poor data provenance, see Hamdy, *supra* note 37.

¹⁰⁷ See Longpre et al., *supra* note 11 (finding that provenance information about training data in many datasets is not full and often inaccurate).

¹⁰⁸ See *id.*

¹⁰⁹ Longpre et al., *supra* note 11, at 13.

¹¹⁰ See Singh, *supra* note 101.

assuming that the use of copyright-protected content in AI training will require authorizations from and/or remuneration for right holders, increased transparency around datasets will mean that AI producers should properly clear rights and, where required, license content before they use the datasets in the training process.¹¹¹ If rights are not properly cleared, right holders are likely to enforce their rights after datasets are disclosed to the public. Rights clearance will further increase the costs of model training.

3. Privacy

As a third challenge, some argue that transparency around training data might lead to privacy violations.¹¹² Training datasets might contain personal data (e.g., photos of individuals, personal information such as their names, age, nationality, and medical information). If AI producers are required to disclose entire datasets, this would mean public disclosure of personal information and potential violation of privacy laws.¹¹³ Indeed, while certain datasets might contain no personal information,¹¹⁴ those that contain such information could not be disclosed to the public due to privacy considerations, unless data subjects have consented

¹¹¹ AI producers are increasingly reaching deals with right holders to use their data for AI training purposes. See AFP, *HarperCollins strikes AI training deal with unnamed company amid rising copyright tensions between publishers and AI firms*, FORTUNE EUROPE (Nov. 22, 2024), <https://fortune.com/europe/2024/11/22/publishing-giants-strike-deals-ai-companies/> [<https://perma.cc/B4JN-R7NQ>].

¹¹² Krook, *supra* note 27, at 14; Yoo, *supra* note 37, at 322; see also Manheim, *supra* note 37; Van Krieken, *supra* note 79; Kohn, *supra* note 76.

¹¹³ Krook, *supra* note 27, at 14; see also Yanisky-Ravid, *supra* note 55; King, *supra* note 75.

¹¹⁴ Those containing non-personal or non-identifiable personal information, such as synthetic data, legal texts, texts or images not associated with specific individuals, and deidentified personal data from medical records.

to such disclosure¹¹⁵ or such disclosure is otherwise allowed under privacy laws.¹¹⁶

In addition, some argue that disclosing datasets might lead to broader security risks, while not specifying the particular risks that would arise.¹¹⁷ It is clear that disclosing the *algorithm* or its parameters might cause security risks. For instance, other persons might modify the parameters to produce outcomes that are against public interests (e.g., hate speech, fake news) or even dangerous (repurposing medical AI to find formulas for chemical weapons).¹¹⁸ However, it is unclear what security risks (apart from privacy ones) could arise out of the publication or disclosure of *training datasets* that were used to train the algorithm.

4. Trade Secrets

Last but not least, releasing datasets or even certain information about datasets might arguably violate AI producers' proprietary interests, and especially trade secrets.¹¹⁹ High quality training datasets are a valuable

¹¹⁵ Clearview AI used personal data (facial images) without authorization from data subjects and was found liable for violations of privacy laws in numerous jurisdictions. *See generally* Kohn, *supra* note 76 (noting that privacy violations occurred due to the collection of data for training purposes).

¹¹⁶ Datasets containing content scraped from social media might contain personal information, and their disclosure might violate privacy laws. For a broader discussion on privacy protection in the AI context, see Manheim, *supra* note 37.

¹¹⁷ Yoo, *supra* note 37, at 322.

¹¹⁸ Rebecca Sohn, *AI Drug Discovery Systems Might Be Repurposed to Make Chemical Weapons, Researchers Warn*, SCIENTIFIC AMERICAN (Apr. 21, 2022), <https://www.scientificamerican.com/article/ai-drug-discovery-systems-might-be-repurposed-to-make-chemical-weapons-researchers-warn/> [<https://perma.cc/W8QA-ZKTM>].

¹¹⁹ *See, e.g.*, Tschider, *supra* note 27, at 711–14 (discussing trade secrets as a challenge for transparency); W. Nicholson Price II & Arti K. Rai, *Clearing Opacity through Machine Learning Essay*, 106 Iowa L. Rev. 775, 790 (2020); *See* Sharon K. Sandeen & Tanya Aplin, *Chapter 24: Trade Secrecy, Factual Secrecy and the Hype Surrounding AI, in*

resource, expensive to collect and prepare, and making these available to the public (and competitors) could undermine the competitiveness of a company. In some cases, AI producers, especially when they work for non-for-profit organizations, release the datasets under open access licenses and do not treat them as proprietary information.¹²⁰ In other cases, AI producers keep them secret. According to a recent study, AI producers are increasingly treating datasets as trade secrets.¹²¹ Even companies that a few years ago were willing to disclose their datasets or information about them, recently stopped releasing such information.¹²² Some AI producers treat not only the data but any information about the training as trade secrets.¹²³ Most jurisdictions around the world grant legal protection to trade secrets, and even recent transparency legislation requires the protection of trade secrets.¹²⁴

Some commentators have argued that it would be unreasonable to request that AI producers disclose

RESEARCH HANDBOOK ON INTELLECTUAL PROPERTY AND ARTIFICIAL INTELLIGENCE 443 (Ryan Abbott ed., 2022).

¹²⁰ There are many open access training datasets available on online repositories such as Common Crawl, RefinedWeb, and the Pile. See Kenny Lee, *Open-Sourced Training Datasets for Large Language Models (LLMs)*, KILI TECHNOLOGY (Jul. 5, 2024), <https://kili-technology.com/large-language-models-llms/9-open-sourced-datasets-for-training-large-language-models> [<https://perma.cc/PET3-EGPT>].

¹²¹ Han, *supra* note 37, at 306; Heldrup, *supra* note 14, at 23.

¹²² *E.g.*, OpenAI initially explained what datasets they used to train initial GPT versions, but for the GPT-4 model, they refused to provide any information about the training process. See Heldrup, *supra* note 14, at 23.

¹²³ *E.g.*, Mistral AI explicitly states that they treat their training data and training approaches as trade secrets. See *Does Mistral AI disclose its training datasets?*, MISTRAL AI, <https://help.mistral.ai/en/articles/156195-does-mistral-ai-communicate-on-the-training-datasets> [<https://perma.cc/5CEW-EKQ4>].

¹²⁴ *E.g.*, EU AI Act, *supra* note 17, at art. 53(1)(b) (“Without prejudice to the need to observe and protect intellectual property rights and confidential business information or trade secrets. . .”).

information that they treat as trade secrets. According to Han, regulations requiring AI producers to unveil their training data would:

[D]isregard the value of creativity of AI producers embodied in training data, which is the secret recipe of their AI. As such, revealing the training data may put AI companies at a great disadvantage. Consequently, developers could lose the incentive to build the creative AI that may overall enhance, rather than replace, the creativity of human creators (footnotes omitted).¹²⁵

On the other hand, it is questionable whether AI producers choose not disclose datasets in order to avoid undermining their competitive position, or for other less legitimate reasons. It is true that there are good reasons not to publish carefully curated high-quality datasets as they have high commercial value.¹²⁶ However, it remains unclear what commercial advantage AI producers gain from non-disclosing which *publicly available* (open) datasets they used. Disclosing the fact that the AI producer used certain publicly available datasets is unlikely to cause them commercial disadvantage as competitors already have access to these datasets and possibly are already using them to pre-train their models.

Instead, a decline in transparency around AI training data in recent years might possibly be associated with multiple legal suits against AI producers claiming violation of intellectual property, privacy, publicity and other laws.¹²⁷ For example, while Stability AI publicly disclosed that they used LIAON5B dataset to train the first versions of the Stable Diffusion algorithm, their transparency around

¹²⁵ Han, *supra* note 37, at 319.

¹²⁶ *Id.*

¹²⁷ See e.g., Knibbs, *supra* note 2 (providing a list of currently pending lawsuits in the US).

training datasets decreased with every subsequent version of the algorithm.¹²⁸ With the latest version of Stable Diffusion algorithm, only vague statements around its training data were issued.¹²⁹

Avoiding legal scrutiny or, at the very least, media coverage and damage to reputation are not the interests that trade secret law is designed to protect. Under international law and many national laws, in order to get trade secret protection for certain information, this information should have an ‘independent economic value,’¹³⁰ or a ‘commercial value.’¹³¹ For instance, under US law, one of the requirements for information to be protected as a trade secret is that “the information derives independent economic value, actual or potential, from not being generally known to, and not being readily ascertainable through proper means by, another person who can obtain economic value from the disclosure or use of the information.”¹³² As Hrdy explains:

[T]rade secret law specifically requires *economic* value. While the term sweeps broadly, recognizing countless ways to capture the value of information, the asserted value must at least be connected to the business or to some form of wealth-seeking activity. The putative economic value cannot stem purely from

¹²⁸ Heldrup, *supra* note 14, at 48.

¹²⁹ *Id.* (“While Stability AI were initially transparent about their training data, they have now chosen to be opaque about what content has been used to train their newer models.”).

¹³⁰ See 18 U.S.C. § 1839(3) (discussing “independent economic value”), see also Defend Trade Secrets Act of 2016, Pub. L. No. 114-153, 130 Stat. 376 (codified as amended in scattered sections of the U.S.C.); Uniform Trade Secrets Act § 1 (amended 1985), 14 U.L.A. 636–37 (2021).

¹³¹ See Agreement on Trade-Related Aspects of Intellectual Property Rights, Apr. 15, 1994

Marrakesh Agreement Establishing the World Trade Organization, Annex 1C, 1869 U.N.T.S. 3 (1994); See *id.* art 39(2)(b) (discussing “commercial value”).

¹³² 18 U.S.C. § 1839(3)(B).

the fact that the secret-holder would prefer that the information be kept confidential or from the fact that disclosure would harm their reputation; the information must be plausibly connected to some underlying economic activity. (footnotes omitted).¹³³

If AI producers keep datasets secret to avoid scrutiny of their questionable AI training practices, it is doubtful whether datasets would meet the independent economic value criteria, since such practices do not provide a direct economic advantage but merely protect them from public or legal scrutiny and from possible reputational damage and legal risks.¹³⁴ If an AI producer is not able to demonstrate that their datasets hold ‘independent economic value.’ they are unlikely to be able to rely on the protection awarded by trade secret laws. In such a case, training datasets that are kept secret by an AI producer would amount to ‘factual’ secrets only. Namely, AI producers could keep them secret, but trade secret law would not protect them from unauthorized disclosure.¹³⁵ Also, if policymakers set requirements to disclose such datasets, this requirement will not conflict with the legal protection for trade secrets.

In summary, the requirement to publish or otherwise disclose training datasets could be a possible option in some cases, especially when AI producers use already publicly available datasets that contain sufficient information to identify specific content and its copyright and licensing status, and do not contain personal or sensitive commercial information. However, mandatory publication of datasets might not be possible if datasets contain personal information or are protected by trade secrets. At the same time, one should be warned that trade secrets cannot be

¹³³ Camilla A. Hrды, *The Value in Secrecy*, 91 *FORDHAM L. REV.* 557, 563 (2022).

¹³⁴ *Id.*

¹³⁵ For a discussion on ‘factual secrets,’ see Sandeen & Aplin, *supra* note 119.

accepted as an excuse to conceal all information about the training data. While trade secrets might be a legitimate reason not to publish entire datasets in some cases, they cannot be used to hide information that does not provide commercial advantage or value for an AI producer, but rather help them avoid unwanted attention, complaints, or legal scrutiny.

B. Summaries of Training Data

Due to challenges and limitations related to mandatory publication of training datasets, some governments have recently proposed that AI producers should provide summaries of their training data instead. This approach addresses some of the challenges related to mandatory publication of datasets, especially those related to privacy and trade secret concerns, but raises doubts as to its effectiveness in ensuring the interests of right holders.

1. Recent Initiatives

Several legislative instruments requiring AI producers to provide summaries of training data have been recently adopted or proposed at the state and federal levels in the US and in Europe.

The California Assembly Bill on Generative AI Training Data Transparency,¹³⁶ which passed in 2024, requires generative AI producers to disclose information about the datasets used to train, test, and validate their AI models. The Bill mandates that documentation includes, among other requirements, a “high-level summary” of the datasets used in the development of the system or service, which will have to be made available on the provider’s website. This high-level summary will include a wide range of information, including information about “the sources or owners of the datasets”; “how the datasets further the

¹³⁶ California Act AB 2013, *supra* note 15.

intended purpose of the artificial intelligence system or service”; “whether the datasets include any data protected by copyright, trademark, or patent, or whether the datasets are entirely in the public domain”; “whether the datasets were purchased or licensed by the developer”, and other information relevant either for copyright holders or other stakeholders (e.g., personal data subjects).¹³⁷ These transparency obligations aim to address a broad range of interests, i.e., not only those of copyright holders but also consumer interests, such as privacy and safety. It applies to generative AI only¹³⁸ and does not apply to non-generative AI models (e.g., those used for object/face identification or predictive models that could be trained on copyright-protected materials).

The US federal government is also considering a few initiatives. The proposed *AI Foundation Model Transparency Act*¹³⁹ calls on the US Federal Trade Commission to create standards for widely used foundation model deployers to publicize certain information about their model, in order to protect the rights of copyright holders and the interests of consumers. Among other types of information, it refers to information about “training data, model documentation, data collection in inference, and operations of foundation models.”¹⁴⁰

Another federal initiative, the proposed *Generative AI Copyright Disclosure Act*,¹⁴¹ would require organizations that operate generative AI systems to submit notice to the US Copyright Office regarding copyrighted works used to

¹³⁷ *Id.* § 3111.

¹³⁸ *Id.* § 3110(c) (“‘Generative artificial intelligence’ means artificial intelligence that can generate derived synthetic content, such as text, images, video, and audio, that emulates the structure and characteristics of the artificial intelligence’s training data.”).

¹³⁹ AI Foundation Model Transparency Act of 2023, *supra* note 16.

¹⁴⁰ *Id.* § 3a(1).

¹⁴¹ Generative AI Copyright Disclosure Act of 2024, H.R. 7913, 118th Cong. (2024).

train the AI system.¹⁴² A person who creates a training dataset is required to submit “a sufficiently detailed summary of any copyrighted works” used in the training dataset and “the URL for such dataset (in the case of a training dataset that is publicly available on the internet at the time the notice is submitted).”¹⁴³

In the EU, the EU AI Act requires 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.”¹⁴⁴ According to the Act, “this summary should be generally comprehensive in its scope instead of technically detailed to facilitate parties with legitimate interests, including copyright holders, to exercise and enforce their rights under Union law.”¹⁴⁵ For example, this could mean “listing the main data collections or sets that went into training the model, such as large private or public databases or data archives, and by providing a narrative explanation about other data sources used.”¹⁴⁶ The EU AI Office is required to provide a template for this summary, which should be “simple, effective, and allow the provider to provide the required summary in narrative form.”¹⁴⁷ Notably, these data

¹⁴² See Miriam Vogel et al., *Is Your Use of AI Violating the Law? An Overview of the Current Legal Landscape*, 26 LEGIS. AND PUB. POL’Y 1029, 1088 (2024).

¹⁴³ Generative AI Copyright Disclosure Act of 2024, *supra* note 141, § 2(a)(1).; *see also*, Content Origin Protection and Integrity from Edited and Deepfaked Media Act (COPIED Act), S.4674, 118th Cong. (2024) (suggesting providing the owners of content the ability attach ownership information to their content and make it unlawful for generative AI models to be trained on or to produce content that has ownership information without the owner’s consent).

¹⁴⁴ EU AI Act, *supra* note 17, art. 53(1)(d).

¹⁴⁵ *Id.* recital 107.

¹⁴⁶ *Id.*

¹⁴⁷ *Id.* *See also id.* art 53(1)(d) (setting a similar requirement). An AI office template is available for download from: *Explanatory Notice and Template for the Public Summary of Training Content for general-*

transparency provisions are intended to address the needs of copyright holders only, while other parts of the EU AI Act, as well as the EU General-Purpose AI Code of Practice, set further transparency duties to address safety, bias and other concerns.¹⁴⁸

Finally, the UK government, in its consultation on Copyright and AI in 2024, has committed to “promoting greater trust and transparency” around AI as one of three main goals.¹⁴⁹ While specific proposals on transparency duties are still to be set, according to the consultation document:

Transparency measures could include requirements for AI firms and others conducting text and data mining to disclose the use of specific works and datasets. Details of web crawlers could also be disclosed, for example including ownership and the purposes for which content is being crawled. They could also include requirements to keep records, to provide certain information on request, or to evidence compliance with rights reservations.¹⁵⁰

General transparency obligations around AI models are also being discussed in many other jurisdictions.¹⁵¹

purpose AI models, EUROPEAN COMMISSION (July 24, 2025), <https://digital-strategy.ec.europa.eu/en/library/explanatory-notice-and-template-public-summary-training-content-general-purpose-ai-models> [<https://perma.cc/2NKQ-BPUQ>].

¹⁴⁸ See EU AI Act, *supra* note 17, art. 50; see also *The General-Purpose AI Code of Practice*, EUROPEAN COMMISSION (July 10, 2025), <https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai> [<https://perma.cc/3Q2H-E83Z>].

¹⁴⁹ UK Copyright and Artificial Intelligence Consultation, *supra* note 7.

¹⁵⁰ *Id.* § 108.

¹⁵¹ See, e.g., Australian Government, *Introducing mandatory guardrails for AI in high-risk settings: proposals paper*, AUSTRALIAN GOVERNMENT DEPARTMENT OF INDUSTRY, SCIENCE AND RESOURCES (2024), <https://consult.industry.gov.au/ai-mandatory-guardrails> [<https://perma.cc/R7RQ-9XKM>]; *Voluntary Code*, *supra* note 19.

Apart from governments around the world, industry has also started developing initiatives and standards to provide summaries of data about AI models and training data. For instance, an AI ‘model card’ is a type of documentation that is created for, and provided with, machine learning models. It functions as a type of data sheet, similar in principle to the consumer safety labels, food nutritional labels, a material safety data sheet or product spec sheets.¹⁵² Standard datasheets, data statements, and data cards provide, in a standardized format, information about AI dataset creators, annotators, language and content composition, innate biases, collection and curation processes, uses, distribution, and maintenance.¹⁵³ Though unevenly adopted, these efforts are widely recognized for improving scientific reproducibility, and responsible AI.¹⁵⁴

One example of such data standardization effort is a Data Nutrition Label.¹⁵⁵ It is based on the US Food and Drug Administration’s Nutrition Facts label and automates data documentation using a registration form, which requires AI producers to provide responses to fifty-five mostly free-text questions about datasets.¹⁵⁶ Another example is the Data & Trust Alliance Data Provenance Standard, which is developed by nineteen corporations (IBM, Nike, Mastercard, Walmart, Pfizer, etc.) and integrates a wide variety of industry data documentation needs into a structured format.¹⁵⁷

All of the above instruments and proposals require the provision of training data summaries; however, there is

¹⁵² See, e.g., Heldrup, *supra* note 14 at 18, 24 (referring to Google Gemma model card and GPT-40 System Card).

¹⁵³ Longpre et al., *supra* note 11, at 19.

¹⁵⁴ *Id.*

¹⁵⁵ *The Data Nutrition Label*, DATA NUTRITION PROJECT, <https://labelmaker.datanutrition.org/> [<https://perma.cc/4H67-7DCP>].

¹⁵⁶ Longpre et al., *supra* note 11, at 19.

¹⁵⁷ *Id.*

a significant divergence as to what goals these proposals seek to achieve (copyright-related only or broader), what information they require to be disclosed, and with relation to which AI models.¹⁵⁸ The analysis below shows that the requirement to provide summaries of training data could indeed address some of the transparency challenges associated with the publication of training datasets; however, there are concerns related to the effectiveness of this approach.

2. Trade Secrets and Privacy

Protection of trade secrets is the most problematic issue when considering the mandatory disclosure of datasets, and thus the publication of the summaries of training data, as an alternative solution, is primarily intended to address this concern. Many of the instruments and proposals on data summaries explained above explicitly commit to protect trade secrets of AI producers and to exclude confidential information from the disclosure requirement.¹⁵⁹ For example, the EU Act recognizes the need to protect trade secrets and confidential business information.¹⁶⁰ The UK government, with relation to its proposals on data transparency, has also committed to protect trade secrets and other confidential information.¹⁶¹

Indeed, by only publishing a summary about the training data used, and not the entire dataset, commercially

¹⁵⁸ See e.g., EU AI Act, *supra* note 17 (discussing ‘foundational models’); California Act AB 2013, *supra* note 15 (discussing ‘generative AI’ models).

¹⁵⁹ See, e.g., EU AI Act, *supra* note 17, art. 53(1)(b) (“Without prejudice to the need to observe and protect intellectual property rights and confidential business information or trade secrets”).

¹⁶⁰ *Id.* (“Without prejudice to the need to observe and protect intellectual property rights and confidential business information or trade secrets”).

¹⁶¹ Intell. Prop. Off., DEP’T SCI., INNOVATION & TECH. & DEP’T CULTURE MEDIA & SPORTS, COPYRIGHT AND AI: CONSULTATION, 2024, CP 1205, ¶¶ 109–10 (UK).

sensitive information would usually be protected. As argued above, even if some AI producers might want to entirely hide any information about training datasets or other training information,¹⁶² such general-level information would not be protected under trade secret law due to a lack of independent economic value,¹⁶³ and thus governments can request its disclosure without interfering with trade secret protection.

Similarly, the requirement to publish summaries of data alone would help avoid any possible privacy risks. Summary descriptions of datasets are unlikely to contain any identifiable personal information, and the information provided is generally unlikely to create other cybersecurity risks.

3. Feasibility and Costs

In terms of technical feasibility and costs, creating summaries, especially if clear instructions and templates are provided, should not cause technical feasibility or unreasonable cost issues.¹⁶⁴ For instance, the EU AI Act requirement to list “main data collections or sets that went into training the model, such as large private or public databases or data archives”¹⁶⁵ would be reasonably easy to satisfy, as would the requirement under the California Generative AI: Training Data Transparency Act to disclose

¹⁶² See, e.g., EUROPEAN COMMISSION, *supra* note 147. Zuzanna Warso & Paul Keller, *Towards Robust Training Data Transparency*, OPEN FUTURE, <https://openfuture.eu/publication/towards-robust-training-data-transparency> (last visited Apr 8, 2025) (proposing a blueprint of the template for the summary of content used to train general-purpose AI models, as prescribed under Article 53(1)d of the EU AI Act).

¹⁶³ Sandeen & Aplin, *supra* note 119, at 455.

¹⁶⁴ See, e.g., EUROPEAN COMMISSION, *supra* note 147. Zuzanna Warso & Paul Keller, *Towards Robust Training Data Transparency*, Open Future, <https://openfuture.eu/publication/towards-robust-training-data-transparency> (last visited Apr 8, 2025) (proposing a blueprint of the template for the summary of content used to train general-purpose AI models, as prescribed under Article 53(1)d of the EU AI Act).

¹⁶⁵ EU AI Act, *supra* note 17, recital 107.

“whether the datasets include any data protected by copyright, trademark, or patent, or whether the datasets are entirely in the public domain.”¹⁶⁶ Meeting these requirements would not be costly or cumbersome, providing that the AI producers have properly documented all necessary information about the training data from the start of the AI project and have it at hand when developing the summary.

Other initiatives, currently in draft form, however, set much higher requirements, and their technical feasibility and cost-effectiveness is still to be determined.¹⁶⁷ For instance, the proposed *AI Foundation Model Transparency Act of 2023* suggests considering for disclosure, “The sources of training data (including, as applicable, personal data collection and information necessary to assist copyright owners or data license holders with enforcing their copyright or data license protections) and whether and how data is collected and retained during inference.”¹⁶⁸ In order to exercise their rights, right holders might need detailed information, including which specific works have been used in a specific AI training project, and their copyright and licensing status. It is questionable how such information can be feasibly provided in a summary description of the training data. As another example, the proposed *Generative AI Copyright Disclosure Act*, would require persons who create datasets to provide a “detailed summary of any copyrighted works used” in a training dataset to the US Copyright Office.¹⁶⁹ What a “detailed summary” would contain, and whether it would be technically feasible, will depend on the exact information that needs to be provided.

¹⁶⁶ CAL. CIV. CODE § 3111(a)(5) (2025); California Act AB 2013, *supra* note 15, § 3111 (a)(5).

¹⁶⁷ *See, e.g.*, H.R. 6881, 118th Cong. § 3(f) (2023).

¹⁶⁸ *Id.*; AI Foundation Model Transparency Act *supra* note 16, § 3(f)

¹⁶⁹ Generative AI Copyright Disclosure Act of 2024, *supra* note 141, § 2(a)(1)(A).

4. Effectiveness

The main challenge is whether the provision of the summary of training data could be effective in ensuring the interests of right holders. It is unclear whether the summary description of datasets could achieve the goal of enabling right holders to license or enforce their rights. The answer will depend on the granularity of information that AI producers will be required to provide in their summaries. If they are required to merely list the names of datasets used and a general description of data (e.g. types of data or URLs), it is doubtful that this information will be sufficient for right holders to identify whether their works were included in the dataset, ask for license, or start an enforcement action. For right exercise purposes they need to know not only whether their works were used but also which works, and from which source these works were taken.¹⁷⁰ They might also need to know when the content was initially collected as the terms of use in different websites might change over time.¹⁷¹ If the right holder is merely told that, for instance, the proprietary dataset contains 11,000 books taken from a specific website,¹⁷² this information does not disclose which books and by which authors are in the dataset and is likely insufficient for individual right holders to exercise their rights. How a summary description of a

¹⁷⁰ Different websites can give access to the same content under different licensing terms; i.e. while some websites might contain properly licensed content, others contain ‘pirated’ content. See Mrva-Montoya, *supra* note 13; Reisner, *supra* note 94; Steven Hawley, *OpenAI Trains Its GPT Model Using Pirated E-Books, Contends Authors’ Lawsuit*, PIRACY MONITOR, (Jun. 30, 2023), <https://piracymonitor.org/chat-gpt-trained-using-pirated-e-books/> [https://perma.cc/35BE-U9KV].

¹⁷¹ See also Heldrup, *supra* note 14, at 3.

¹⁷² See, e.g., Richard Lea, *Google Swallows 11,000 Novels to Improve AI’s Conversation*, The Guardian, (Sep. 28, 2016), <https://www.theguardian.com/books/2016/sep/28/google-swallows-11000-novels-to-improve-ais-conversation> [https://perma.cc/UG44-QQXD].

training dataset can meet such specific information needs of right holders remain unclear.

On the other hand, the summary description might be more effective if, in addition to it, an AI producer also lists further sources where more detailed information could be found. For instance, with relation to Phi-1 and Phi-2 models, Microsoft disclosed titles of datasets and references to papers by third parties with more details about these models, including the dataset “SlimPajama” used in training Phi-2.¹⁷³ This dataset is publicly available and right holders can download and then search it for specific content.¹⁷⁴ On the other hand, it is a cumbersome process requiring specific technical skills,¹⁷⁵ and it is questionable which of the right holders could afford such effort and whether it is reasonable to expect them to fund the search.

One last question is how and who will ensure that the summary statements are correct and not missing any information. In order to ensure the quality of such summary descriptions, auditing or other complementary measures would have to be created, and responsible authorities would have to be assigned that would verify the correctness of the summaries. Such verification or audit could be funded by government¹⁷⁶ or by AI producers required to conduct such

¹⁷³ See, e.g., *Microsoft/Phi-2 Model Card*, HUGGING FACE, <https://huggingface.co/microsoft/phi-2> [<https://perma.cc/KN7D-T4HP>] (last visited Apr. 14, 2025); Suriya Gunasekar et al., *Textbooks Are All You Need*, ARXIV (Oct. 2, 2023), <http://arxiv.org/abs/2306.11644> [<https://perma.cc/J8RQ-A9MF>].

¹⁷⁴ HUGGING FACE, *supra* note 173.

¹⁷⁵ See Heldrup, *supra* note 14, at 25–32 (describing how in order to understand what data Microsoft’s Phi models were trained on, researchers had to check numerous sources and found limited information).

¹⁷⁶ This has been done before, as compliance with transparency obligations under the Artificial Intelligence Act is supervised by the EU AI Office. See EU AI Act, *supra* note 17; see generally Busuioc et al., *supra* note 17 (discussing the EU AI Act).

an audit. In either case this would add to the costs of ensuring data transparency.

Overall, while a requirement to publish summaries of training data is likely to address trade secret and privacy concerns, and probably will not be problematic from technical feasibility and cost perspectives, the effectiveness of this measure in addressing the needs of right holders is questionable. Unless combined with other measures (e.g. links to full datasets), this approach is unlikely to result in the provision of information that is sufficient to enable right holders to exercise their rights.

C. Certification of Datasets

A third alternative solution is a certification system which would allow database creators and/or AI system producers to acquire certification that their datasets are ‘legal’ from a copyright perspective, i.e. the rights of all content items have been properly cleared. Certification of data for *privacy and bias* reasons has been proposed on a number of occasions both in literature¹⁷⁷ and by law makers,¹⁷⁸ and a certification system confirming their legality from *copyright* perspective has been recently

¹⁷⁷ See Yanisky-Ravid & Hallisey, *supra* note 55 (suggesting auditing systems for AI bias); Greg Satell & Josh Sutton, We Need AI That Is Explainable, Auditable, and Transparent, Harvard Business Review (Oct. 28, 2019), <https://hbr.org/2019/10/we-need-ai-that-is-explainable-auditable-and-transparent> [<https://perma.cc/345Q-4KQ4>] (suggesting auditing systems for AI bias).

¹⁷⁸ For example, in 2017, New York City passed an algorithmic transparency law. See N.Y.C., N.Y., LOCAL LAW 49 (Jan. 17, 2018) (requiring a creation of a task force to provide recommendations on certain AI systems used by New York City agencies to mitigate potential discriminatory impacts caused by such systems). See also Algorithmic Accountability Act, H.R. 2231, 116th Cong. (2019) (proposing that covered companies subject to the bill must evaluate training data of their AI systems to determine the systems’ “impacts on accuracy, fairness, bias, discrimination, privacy, and security”).

discussed too.¹⁷⁹ According to Han's proposal, the AI producer would disclose the training data to a designated certification body which would assess whether copyright of all used data has been properly cleared.¹⁸⁰ While this solution can indeed address trade secret and privacy issues, its effectiveness, feasibility and costs, especially for large and complex datasets, raise concerns.

1. Trade Secrets and Privacy

According to this approach, AI producers will not need to disclose any information to the public but would have to disclose all information about their datasets to certifying bodies. Assuming that these certifying bodies would be subject to confidentiality duties, trade secrets of AI producers would not be affected.¹⁸¹ Indeed, well established certification bodies, such as the Food and Drug Administration which certifies food and therapeutic goods, are subject to strict confidentiality duties.¹⁸²

Similarly, such a certification system would arguably address privacy concerns that might arise when information is being disclosed to the public. A certification scheme could allow sharing datasets containing personal information with appropriate certification bodies, under strict non-disclosure requirements.¹⁸³

On the other hand, there might be interest among stakeholders or the general public (especially researchers or advocacy organizations) to obtain access to at least summary certification reports, which will have to disclose at least some information. These reports, however, would be

¹⁷⁹ See Han, *supra* note 37, at 321.

¹⁸⁰ *Id.*

¹⁸¹ *Id.* at 315

¹⁸² U.S. FOOD & DRUG ADMIN., REGULATORY PROCEDURES MANUAL § 3-6 (2024).

¹⁸³ *Id.*

unlikely to disclose information that would constitute commercial value or be of personal nature.¹⁸⁴

2. Feasibility, Cost and Effectiveness

The aim of the certification system would be to provide assurances to right holders that the rights are properly cleared and there is no need of further licensing or enforcement action, which would arguably satisfy the interest of right holders.¹⁸⁵ Certification schemes have been successful in other areas of law¹⁸⁶ and this solution indeed might be possible with smaller or less complex datasets. For instance, it might be possible to audit and certify a dataset which contains content that is clearly not protected by copyright,¹⁸⁷ or content to which the rights have been clearly licensed.¹⁸⁸

In many cases, however, feasibility of such certification is questionable. Firstly, AI producers should have their data collection and preparation well-documented.

¹⁸⁴ Compare *id.*, with 510(k) Premarket Notification, U.S. FOOD & DRUG ADMIN.,

<https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfPMN/pmnm.cfm> (last visited Mar. 1, 2026) (on file with author) (providing a database of summaries with FDA summaries of medical devices via 510(k) pre-market notification pathway).

¹⁸⁵ See Han, *supra* note 37.

¹⁸⁶ See Yanisky-Ravid & Hallisey, *supra* note 55, at 476.

¹⁸⁷ For example, datasets containing medical images.

¹⁸⁸ For example, content taken from websites that explicitly grant a (free or paid) license to use content for training purposes; or when an individual licensing agreement is signed with the right holder. See AFP, *supra* note 111; Mackenzie Ferguson, Publishers and AI Companies Strike Groundbreaking Deals in 2024: A New Era for Media Partnerships!, OPENTOOLS (Dec. 27, 2024), <https://opentools.ai/news/publishers-and-ai-companies-strike-groundbreaking-deals-in-2024-a-new-era-for-media-partnerships> [https://perma.cc/Q2HC-39KH] (providing examples of when an individual licensing agreement is signed with the right holder). Another example is content taken from websites that explicitly grant a free or paid license to use content for training purposes.

As the standards are only emerging in data management, including data documentation,¹⁸⁹ it is likely that many companies might not yet have appropriate records of their training data. Second, keeping in mind the fact that a large amount of content contained in large datasets might reach billions,¹⁹⁰ certifying that rights to each piece of content have been properly cleared becomes essentially impossible. Most of the data items in such datasets are likely not to have information about their licensing status. Research shows that even when ownership and/or licensing status is listed in certain public data libraries, in many instances the information is not full or is incorrect.¹⁹¹

Further, identifying the copyright status of each content item will be very difficult, if not impossible, even for certifying bodies. With relation to content with limited metadata, it will not be clear who the author of a specific content item is, and whether and when they have passed away, which would make it impossible to determine whether the piece is still protected by copyright, or in some cases, whether it is protected at all.¹⁹² Even if the author is

¹⁸⁹ See EU AI Act, *supra* note 17, art. 2(d), art. 11(1); EUROPEAN COMMISSION, *supra* note 148 (setting data documentation requirements in the EU); see NIST's AI Standards "Zero Drafts" Pilot Project to Accelerate Standardization, Broaden Input, NIST (Sept. 12, 2025), <https://www.nist.gov/artificial-intelligence/ai-research/nists-ai-standards-zero-drafts-pilot-project-accelerate> [<https://perma.cc/6FX8-YWCE>] (providing an example of a U.S. initiative).

¹⁹⁰ See, e.g., Admin Staff, *How Was Stable Diffusion Trained?*, NightCafe (Oct. 9, 2022), <https://nightcafe.studio/blogs/info/how-was-stable-diffusion-trained> (explaining how Stable Diffusion was trained on LAION-5B's full collection of 5.65 billion image-text pairs).

¹⁹¹ Longpre et al., *supra* note 39.

¹⁹² For example, if the work was created with the assistance of AI, determining its copyright status will be very complex. See Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence, 88 Fed. Reg. 16190 (Mar. 16, 2023) (demonstrating that if a work was created with the assistance of AI, determining its copyright status will be very complex).

identified, clarifying the status of the rights in the work might be impossible as even the certifying authority might not have access to information on whether and to whom the author has transferred the rights. An ‘orphan work’ problem that was raised by libraries in digitization projects a few decades ago¹⁹³ would arise again here, but on a much larger scale.

Also, since different national laws have different scopes of rights and exceptions with relation to the use of content in the AI context,¹⁹⁴ the jurisdictional problem would arise—which laws should be relied upon for the rights clearance purpose? It is often assumed that AI producers are subject only to the laws of the country where the dataset is developed and where the training is taking place. Thus, if an AI model was trained in the US, it has been assumed that US copyright law will apply.¹⁹⁵ However, the EU AI Act suggests that AI models offered inside the EU should comply with EU copyright law, even if they are trained overseas.¹⁹⁶ Other jurisdictions are considering similar approaches,¹⁹⁷ which suggests that clearing rights under the laws of the ‘training country’ might be insufficient if the AI system is to be marketed internationally.

Keeping in mind the complexity of checking the copyright and licensing status, there would be significant costs involved in such a procedure.¹⁹⁸ The larger and the

¹⁹³ See generally David R. Hansen et al., *Solving the Orphan Works Problem for the United States*, 37 THE COLUMBIA J.L. & ARTS 1 (2013) (providing a general discussion of the orphan works problem).

¹⁹⁴ See *supra* Section III.C.2 (discussing references to different copyright exceptions under the US, UK, EU and Australian laws).

¹⁹⁵ See, e.g., Shen, *supra* note 10; Maffioli, *supra* note 8; Sag et al., *supra* note 8 (examining the applicability of US fair use doctrine to US-developed models).

¹⁹⁶ EU AI Act, *supra* note 17, art. 53(1)(c).

¹⁹⁷ See *UK Copyright and Artificial Intelligence Consultation*, *supra* note 7.

¹⁹⁸ See also Han, *supra* note 37, at 314.

more complicated the dataset, the more costly its certification would be. The price will have to be paid by AI producers who are likely to transfer the costs onto AI users. In addition, the certification price is separate from the price that AI producers will have to pay to get licenses of content protected by copyright, and the more content is used, the costs associated with clearing the rights, including the licensing fees, will be higher.

Considering the legal complexity of clearing the rights and the many laws that possibly apply for AI models to be used at an international scale, it is questionable whether right holders could and would trust the certification outcomes. Some questions—e.g. whether specific content is protected under copyright, whether a specific copyright exception applies to a given scenario or not—might be contested, and could only be resolved with finality by courts alone. Leaving the resolution of such questions to a certification body might not satisfy the right holders or the public interest.

In addition, even if a dataset certification system is established by law and publication of detailed information about datasets is not required, right holders might have a legitimate interest in knowing whether their content was used in the training of AI, and in which particular AI systems. This might be especially relevant in jurisdictions which provide authors with strong moral rights, such as authorship and integrity rights.¹⁹⁹ The authorship right requires the identification of the author whose work has been used, while the integrity right enables authors to prohibit the use of their work if this negatively affects author's honor or reputation. Even if a certification body confirmed that

¹⁹⁹ See Berne Convention for the Protection of Literary and Artistic Works, art. 6bis, Sept. 9, 1886, S. Treaty Doc. No. 99-27 (1986); see also Adolf Dietz, *The Moral Right of the Author: Moral Rights and the Civil Law Countries*, 19 Colum.-VLA J.L. & Arts 199 (1994) (explaining moral rights in civil law countries).

copyright has been cleared, right holders might demand the disclosure of some information based on moral rights.²⁰⁰ The US protection of moral rights is limited²⁰¹ and thus this problem is unlikely to arise with relation to AI systems trained in the US and intended for the US market, but would be relevant for models to be offered in the international market.

IV. A PROPOSED APPROACH TO AI DATA TRANSPARENCY

The above analysis has demonstrated that none of the currently available or proposed solutions—the mandatory publication of datasets, summaries of training data or the certification of datasets—are suitable in both ensuring sufficient information about datasets for right holders and taking into consideration the interests of AI producers and public members. While the publication of datasets might enable right holders to identify whether their works have been used in AI training and is generally possible from a technical and cost perspective, it might not disclose all information needed by right holders; in some cases, it is not viable due to significant risks to privacy and trade secrets. Publishing summaries of training data has a potential to address the latter issues but their effectiveness in satisfying the interests of right holders is questionable: will they provide information that is sufficient for right holders to exercise their rights? Finally, a certification system for

²⁰⁰ See, e.g., RITA MATULIONYTE, CAN AI INFRINGE MORAL RIGHTS OF AUTHORS AND SHOULD WE DO ANYTHING ABOUT IT: AN AUSTRALIAN PERSPECTIVE (2023), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4016001 (on file with author) (discussing the application of moral rights in the AI context in Australia).

²⁰¹ See, e.g., Cyrill P. Rigamonti, *Deconstructing Moral Rights*, 47 HARV. INT'L L.J. 353 (2006); Gerald Dworkin, *The Moral Right of the Author: Moral Rights and the Common Law Countries*, 19 COLUM.-VLA J.L. & ARTS 229 (1994).

datasets could potentially address privacy and trade secrets concerns, and is meant to provide reassurances to right holders that the rights to all content have been properly cleared. While this solution might be workable for smaller datasets, in case of large datasets with limited metadata, it might be unfeasible to conduct certification for thousands or billions of content items and the associated costs would be unbearable.

In order to address the limitations of current approaches to data transparency, this paper calls for a new—flexible, accountable and holistic (FAH)—approach to AI data transparency.

A. *The FAH Approach*

1. Flexible

The first key feature of the proposed approach is its *flexibility*. This approach suggests imposing on AI producers a general flexible obligation to ensure effective transparency around AI training data, together with a ‘bucket-list’ of data transparency measures from which AI producers can choose the ones that are most suitable for a specific dataset or a specific AI project. This bucket-list of measures, which would preferably be implemented in an industry code of best practices²⁰² or an industry standard,²⁰³ would explain, among other things, which specific data transparency measures are most suitable for different types of datasets and AI projects, considering effectiveness, feasibility, cost, trade secrecy and privacy perspectives.²⁰⁴

²⁰² See, e.g., EUROPEAN COMMISSION, *supra* note 148 (setting data documentation requirements in the EU); *Voluntary Code*, *supra* note 19. Cf. EUROPEAN COMMISSION, *supra* note 147.

²⁰³ See AI Foundation Model Transparency Act, *supra* note 16.

²⁰⁴ For a discussion of the strengths and weaknesses of industry code as a regulatory measure, see Neil Gunningham & Joseph Rees, *Industry Self-Regulation: An Institutional Perspective*, 19 L. & POL’Y 363 (1997).

The approach is flexible as AI producers could choose from several transparency approaches, listed in the code of best practices (e.g. disclosure of a dataset, publication of a detailed summary of datasets, a certification of dataset or others). They could even be allowed to adopt a novel approach to ensuring data transparency, as long as it is *effective*, i.e. it leads to the disclosure of information that is sufficient for right holders to exercise their rights.²⁰⁵

2. Accountable

The second feature of this approach is a clear distribution of *accountability*. Under this approach, transparency obligations would be distributed among relevant actors in the AI supply chain.²⁰⁶ The code of best practice would set specific transparency-related obligations to different AI actors, depending on their role in the AI supply chain. Organizations which create datasets (dataset creators) should be responsible for providing dataset descriptions (e.g. dataset cards) and relevant metadata for each content item included in the dataset; the code would have to set guidelines, or refer to specific industry standards, on what metadata is to be provided and how.²⁰⁷ In order to address right holder needs, each content item might have a data card indicating a source from where the content item was taken (URL or otherwise), authors and/or owners of content, copyright and licensing status, as far as the collection and integration of this information is possible and

²⁰⁵ See generally *supra* Section III.C.3 (providing a discussion on the effectiveness of allowing the adoption of novel approaches to ensuring data transparency).

²⁰⁶ See generally EU AI Act, *supra* note 17, art. 53 (setting obligations on providers of general-purpose AI models only).

²⁰⁷ For examples of dataset cards see DATA NUTRITION PROJECT, *supra* note 155 (providing examples of dataset cards that do not consider copyright holder interests).

reasonable.²⁰⁸ The organization conducting AI training (AI training organization) should be responsible for the documentation of the training process, including the documentation of datasets that were used to train (pre-train or finetune) the model, and how it was processed during the training process.²⁰⁹ This would allow right holders to identify which of their works were used when training a specific system, and how they were modified during the process.

AI dataset creators and AI training organizations would be required to share the relevant information with organizations further down the AI supply chain,²¹⁰ and in particular, with AI providers who offer the AI system in the market and who will be directly responsible for providing relevant stakeholders (right holders, general public, certification bodies) with sufficient information about the training data.²¹¹

²⁰⁸ See e.g., id. (giving examples of industry standards for data cards); Mahima Pushkarna, Andrew Zaldivar & Oddur Kjartansson, *Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI*, in FACCT '22: PROCEEDINGS OF THE 2022 ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 1776 (2022), <https://dl.acm.org/doi/10.1145/3531146.3533231> (on file with author) (providing examples of industry standards for data cards)..

²⁰⁹ See EUROPEAN COMMISSION, *supra* note 147 (providing a template of what information should be provided about the AI training process).

²¹⁰ See generally KATHERINE LEE ET AL., TALKIN' 'BOUT AI GENERATION: COPYRIGHT AND THE GENERATIVE-AI SUPPLY CHAIN (2023), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4523551 [<https://perma.cc/UMK9-LER4>] (providing a general discussion of AI supply chain).

²¹¹ See HAROLD BOOTH ET AL., U.S. DEP'T OF COMMERCE, SECURE SOFTWARE DEVELOPMENT PRACTICES FOR GENERATIVE AI AND DUAL-USE FOUNDATION MODELS (2024), <https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-218A.pdf> [<https://perma.cc/3EV8-46AZ>] (setting documentation and disclosure requirements through NIST); EUROPEAN COMMISSION, *supra* note 148 (setting detailed documentation requirements).

Such actor-specific duties would clarify the accountability of all AI actors in achieving effective data transparency. It will ensure that necessary information is documented from initial stages of the project and, when the AI system is offered in the market, this information is made available to relevant stakeholders (e.g. right holders, government or certifying bodies).

3. Holistic

The third feature of this proposal is its holistic approach to AI data transparency. The previous analysis has made it clear that data transparency rules cannot be set alone disregarding the existing copyright law framework.²¹² The justification and effectiveness of data transparency rules are highly dependent on other elements of the copyright law framework, and thus, the transparency rules will have to be a part of a broader and holistic revision of copyright law. For instance, it is not possible to require data transparency if copyright law does not provide right holders with at least certain rights with relation to the use of their content in the AI development process.²¹³ Also, it is not appropriate to require AI producers to disclose detailed information about datasets if this creates unmanageable legal risks for them. For instance, if ensuring data transparency means that AI producers will be exposed to copyright lawsuits while simultaneously the copyright law framework does not provide any means for AI producers to secure necessary copyright licenses and comply with law,²¹⁴ imposing

²¹² See also Adam Buick, *Copyright and AI Training Data—Transparency to the Rescue?*, 20 J. INTELL. PROP. L. & PRAC. 182 (2025) (arguing that data transparency alone cannot fix the problems of copyright law in the AI context).

²¹³ See Sag et al., *supra* note 63 (arguing that transparency is not needed if the use of content in AI training amounts to fair use).

²¹⁴ See Matulionyte, *supra* note 53 (emphasizing that current licensing frameworks do not allow AI producers to get one-stop-shop licenses for large amounts of content required for large AI training projects).

transparency duties on AI producers would be an unfair and imbalanced solution.

To address this complex challenge, it is suggested that a holistic perspective to AI data transparency should be taken. In parallel to the general flexible obligation to ensure effective data transparency, governments should consider at least three additional copyright law measures that are closely intertwined with AI data transparency rules.²¹⁵

First, before data transparency duties are established, law makers need to clarify the scope of exclusive rights and copyright exceptions, as they apply to the AI development process. The scope of rights, as they apply in the AI training context, will determine the scope of transparency duties. If certain uses of content in AI training are exempted from the scope of copyright law,²¹⁶ then transparency obligations should not extend to them, as there is no justification to require transparency around uses that are not covered under exclusive rights.²¹⁷

Second, law makers are invited to consider a rights clearing mechanism that would allow AI producers to license rights in the content that they need for training purposes.²¹⁸ While the specific parameters of such a rights clearance mechanism requires a separate analysis,²¹⁹ Pits overall purpose would be to enable AI producers to obtain necessary licenses for content protected under copyright, without unreasonably high transaction costs. Information

²¹⁵ See also Han, *supra* note 37 (suggesting several interrelated measures to solve data transparency problems).

²¹⁶ For example, in the EU, the use of content in non-commercial TDM projects are covered by exceptions. See EU DSM Directive, *supra* note 48, arts. 3–4.

²¹⁷ Sag et al., *supra* note 63, at 33–34 (arguing that transparency is not needed if the use of content in AI training amounts to fair use).

²¹⁸ See Han, *supra* note 37, at 322 (suggesting “Online Data Collection License” as an option).

²¹⁹ For a general discussion on remuneration for the use of works in AI training, see Senftleben, *supra* note 10; Shen, *supra* note 10.

about licensing conditions could then be listed in dataset cards and shared with right holders as a part of a data transparency framework. This would address the problem that currently most datasets face, namely, the absence of reliable information about copyright and licensing status.²²⁰ Also, the licensing scheme would minimize or eliminate legal risks for AI producers to be sued for copyright infringement. If AI producers are not concerned about possible legal suits, they would be more willing to share information about AI training datasets, and effective data transparency would be easier to achieve.

Finally, law makers are invited to consider a limited liability regime that would benefit AI producers complying with rights clearance and data transparency duties. If the AI producer follows best reasonable efforts to clear the rights and complies with data transparency duties, their liability for any resulting infringement could be limited.²²¹ For example, if despite reasonable efforts by the AI producer, certain rights have not been cleared or certain information about training data is not made transparent, an AI producer would not be required to pay damages, but might still be subject to other remedies. They could be required, when notified, to remove certain content from the dataset and/or pay licensing fees for the content that was used in the training and which cannot be removed from the already trained AI model.²²² This limited-liability regime would be in principle similar to the limited-liability, or ‘safe harbor’, regime that was

²²⁰ See the discussion *supra* note 219. See also Longpre et al., *supra* note 39.

²²¹ See Yanisky-Ravid & Hallisey, *supra* note 55 (suggesting a safe harbor regime for AI producers, but in the bias context).

²²² While ‘unlearning’ data by AI models is difficult, new approaches are being developed to address this. See, Ben Wodecky, AI Models Can Now Selectively ‘Forget’ Data After Training, AI Business (Oct. 6, 2023), <https://aibusiness.com/ml/ai-models-can-now-selectively-forget-data-after-training> (on file with author).

introduced for internet service providers in earlier years of the Internet.²²³

While the scope of exclusive rights in the AI context, and specific parameters of a licensing scheme and limited liability regime would require a separate study, they would be necessary elements of a holistic copyright framework, where data transparency duties are balanced with measures that protect the legitimate interests of AI producers. This would lead to effective data transparency and a well-balanced copyright law framework.

B. The FAH Approach and Current Challenges to AI Data Transparency

In comparison to current data transparency measures discussed above, the FAH data transparency framework is also more likely to address the challenges that the existing approaches to data transparency face. It is likely to be more effective as it allows for choosing data transparency measures that are more effective and suitable to a specific dataset. It is also likely to ensure that the interests of AI producers, such as trade secrets, financial and legal risk considerations, and the public interests, including privacy, are properly taken into consideration.

1. Effectiveness

The analysis of current measures has shown that they all are facing certain challenges to their effectiveness. The publication of datasets is not an effective measure to ensure

²²³ See 17 U.S.C. § 512 (listing safe harbor provisions in U.S. law); Council Directive 2000/31, 2000 O.J. (L 178) 1, 4, 7. See generally Niva Elkin-Koren, Yifat Nahmias & Maayan Perel, Is It Time to Abolish Safe Harbor? When Rhetoric Clouds Policy Goals, 31 *Stan. L. & Pol’y Rev.* 1 (2020) (discussing the safe harbor act in relation to copyright law); Matthew Sag, Internet Safe Harbors and the Transformation of Copyright Law, 93 *Notre Dame L. Rev.* 499 (2017) (evaluating the strengths and weaknesses of safe harbor rules under copyright law).

right holders' interests if datasets, even if publicly available, do not have adequate metadata, including information about copyright, ownership and licensing status of content items included in the dataset.²²⁴ Data certification is not going to be effective if it is not legally possible or feasible to certify the copyright and licensing status of each data item in the training dataset.²²⁵ Publication of data summaries will not be effective if they do not disclose detailed information that right holders require for rights' exercise purposes.²²⁶

The proposed FAH approach aims to address these challenges to effectiveness. First, the list of data transparency measures, to be incorporated in a code of best practices, will be developed by all stakeholders who will agree on measures that are sufficiently effective, i.e. result in information that meet right holders' transparency needs. The code would have to identify and describe different data transparency measures that are most effective for different datasets or different AI projects. For instance, if a dataset does not contain personal information or trade secrets, the most effective measure could be the publication of the dataset, as long as it has adequate data provenance and licensing information. In contrast, when a dataset contains personal information or trade secrets, other transparency measures would be most suitable, such as data certification confirming the compliance with copyright law, or a summary of data and an ability of right holders to request further information about specific works used in the training process. The code of best practices would have to consider a variety of approaches that are effective in various scenarios. It would consider the most recent technological developments, and would be revised over time to reflect the new technical and legal developments, to ensure that most up-to-date measures are listed.

²²⁴ See *supra* Section III.A.1.

²²⁵ See *supra* Section III.C.2.

²²⁶ See *supra* Section III.B.4.

This approach is also likely to be more effective than previous ones as all actors in the AI supply chain would have clearly defined duties in ensuring data transparency.²²⁷ Transparency duties cannot be attached to AI providers only. AI providers offering AI systems to the market will not be able to ensure data transparency if data provenance and licensing information was not properly recorded when datasets were created, and the training process was not well documented.²²⁸ Clear allocation of transparency-related duties to all actors in the AI supply chain, including for organizations creating the datasets and training the AI algorithms, is essential to ensure effectiveness of data transparency rules.

Finally, it is suggested that right holders are entitled to raise complaints with a competent authority or bring a legal action if they believe that the data transparency measures adopted by an AI producer do not meet the effectiveness standard. This would discourage AI providers from adopting data transparency measures that are not effective from a right holder's perspective, such as publication of dataset summaries that do not contain information sufficient for copyright holders to exercise their rights.²²⁹

2. Feasibility and Costs

The proposed approach will increase feasibility and address some of the cost and legal risks related concerns by AI producers. The above analysis has shown that while the publication of a dataset or certification of datasets are measures with certain strengths, it might not be technically

²²⁷ See *supra* Section IV.A.2.

²²⁸ Cf. EU AI Act, *supra* note 17, art. 53 (requiring providers of general-purpose models to “draw up and keep up-to-date the technical documentation of the model, including its training and testing process and the results of its evaluation”).

²²⁹ See *supra* Section III.B.4.

feasible or too costly to collect and include in the dataset detailed information that right holders need for rights exercise purposes, or that certification bodies need to confirm that the dataset complies with obligations under copyright law.²³⁰ The FAH approach to data transparency will help address this challenge in a few ways.

First, as already mentioned, the code of best practice would set, among other things, requirements for dataset creators on how to record data provenance in the datasets, and standards for AI training organizations on how to document the training process, as well as duties to share this information with AI actors down the supply chain.²³¹ This will ensure that AI providers offering AI systems in the market have the required information at hand and can make it available to relevant stakeholders (public, right holders or certification bodies).

Second, the feasibility of providing *licensing* information about AI datasets will improve when AI producers have a mechanism to license the rights, as proposed under the FAH approach.²³² If an AI producer gets a license for the content, incorporating licensing information in a dataset card or in metadata for each item will be a feasible task.

In addition, if data transparency duties for all AI actors are established and licensing schemes are in place, then the certification of datasets would become a more feasible measure. When datasets contain information about copyright and licensing status, it would be more feasible for certification bodies to assess whether the rights to the content have been properly cleared, and certify that the data was used in compliance with copyright laws. Thus, data certification might become a feasible option not only for

²³⁰ See *supra* Sections III.C.2, III.A.2.

²³¹ See *supra* Section IV.A.2.

²³² See *supra* Section IV.B.2.

small or ‘simple’ datasets, but also for larger and more complicated ones.

In addition, a rights licensing mechanism, together with a proposed limited liability regime, would also protect AI producers from unreasonable legal risks. As long as they follow the best data transparency practices and properly clear and license rights (e.g. via a rights clearance mechanism proposed under the FAH framework), they will be protected from unreasonable legal risks, especially from statutory damages.²³³

In terms of costs, while transparency measures themselves (e.g. publication of datasets, data summaries) are not likely to attract significant costs, collecting data provenance information, and especially clearing/licensing rights, would certainly result in higher costs for AI producers. On the other hand, increased costs are reasonable keeping in mind the benefits AI producers would receive (limited liability) and are indispensable to balancing the interest of all stakeholders involved, especially with the interests of right holders whose works are used to train AI models.

3. Protection of Trade Secrets and Privacy

Finally, the FAH approach would help address challenges related to trade secrets and privacy as it allows AI producers to choose data transparency measures that are most suitable with relation to a specific dataset. For example, if a dataset is not protected as a trade secret and does not

²³³ Statutory damages in copyright cases in the US can be very significant. *See, e.g.*, Adam Philipp, Jury Awards Record Damages in AI-Assisted Copyright Infringement Case, *AEON Law* (Jan. 8, 2024), <https://aeonlaw.com/jury-awards-record-damages-in-ai-assisted-copyright-infringement-case/> [<https://perma.cc/T95C-ZA9J>]; *see also* Oren Bracha & Talha Syed, The Wrongs of Copyright’s Statutory Damages, 98 *Tex. L. Rev.* 1219 (2019) (providing criticism of statutory damages under copyright law).

contain personal information, an AI producer might choose to publish the training dataset in their entirety, or disclose the name and location of an already published dataset.²³⁴ This option will allow right holders to search and identify whether their works form a part of this dataset and have been used in the training of a particular AI system.

When an AI producer uses datasets that contain personal information or are protected as trade secrets,²³⁵ AI producers will not be required to disclose them in full but can choose other data transparency measures provided in the code of best practices. For instance, they might be able to provide the names of datasets, coupled with relevant industry certification which confirms that the rights in the content contained in those datasets are properly cleared.²³⁶ Importantly, the proposed framework would only protect trade secrets over information that provides AI producer with commercial advantage (i.e. those protected under law), and not information that is kept secret to avoid scrutiny (factual secrets).²³⁷

Finally, apart from the measures in the code of best practice, AI producers should be allowed to adopt new, innovative approaches to ensuring transparency around training data, as long as they are both effective and avoid disclosing personal information or trade secrets. For instance, AI producers might develop a mechanism which

²³⁴ Some examples of data libraries containing multiple open datasets include: Kaggle, UCI Machine Learning Repository, and Google Dataset. Search are data libraries containing multiple open datasets, *see* SMARTONE.AI, *supra* note 88.

²³⁵ *See* Heldrup, *supra* note 14 (arguing that most datasets are currently claimed as trade secrets by AI producers).

²³⁶ Where relevant, certification bodies could certify that data was collected in compliance with privacy laws and non-discrimination standards. *See*, e.g., Yanisky-Ravid & Hallisey, *supra* note 55 (suggesting a certification system to manage AI bias).

²³⁷ *See supra* Section III.A.4; *see also* Sandeen & Aplin, *supra* note 119 (discussing the difference between trade secrets and factual secrets).

allows them to provide necessary information upon request from the right holder.²³⁸ Such an approach might be especially relevant when the publication of entire datasets is not feasible or possible (e.g. for privacy, trade secret or other safety reasons). Alternatively, an AI producer might provide a system for right holders to submit the works that should be excluded from AI training, and then remove these works from their training datasets.²³⁹

V. CONCLUSION

Transparency is a long-standing concept and, in the AI context, its importance for different stakeholders is well established. AI *data* transparency, as a part of broader AI transparency concept, has become urgently needed in the copyright space to enable copyright holders to exercise their rights with relation to their content used in AI training. Ensuring AI data transparency, however, is not straightforward. The current and recently proposed measures, such as a mandatory public disclosure of datasets, publication of summaries of datasets or data certification measures, are with clear limitations, and would not be suitable for all AI datasets and projects, as they may raise privacy, confidentiality, effectiveness, technical feasibility, and cost challenges. This paper thus proposes a new framework for AI data transparency in copyright context, which is flexible, accountable and holistic (FAH

²³⁸ See *UK Copyright and Artificial Intelligence Consultation*, *supra* note 7 (suggesting that requests of information could be one measure). One automated option is to allow right holders search if their content is a part of the dataset. See *SPAWNING AI*, *supra* note 93.

²³⁹ This approach should be compared to the Youtube Content ID program which allows right holders to submit their content for the purpose of identifying illegal content on the Youtube platform. See *How Content ID Works*, YOUTUBE HELP, <https://support.google.com/youtube/answer/2797370?hl=en> [<https://perma.cc/W87G-PUUB>] (last visited Mar. 1, 2025).

framework). It sets a flexible duty on AI producers to ensure effective transparency around AI training data which allows them choose transparency measures that are most effective and suitable for a given dataset. It ensures accountability by requiring specific transparency-related duties from different AI actors in the AI supply chain. Finally, to address complex feasibility, legal risk and cost challenges, it adopts a holistic approach to data transparency rules and suggests that they should be combined with other copyright law measures, such as a licensing scheme and a limited liability regime, to ensure both the effectiveness of the data transparency rules and a properly balanced copyright law framework. The FAH framework is expected to address the challenges that current data transparency measures face: it is both likely to result in information that right holders need, and adequately take into consideration the interest of AI producers and the public (users and data subjects).