NETWORK LAW REVIEW

ARTICLE

Discit ergo est: Training Data Provenance And Fair Use

Robert Mahari and Shayne Longpre

***

1. Introduction

AI is the product of its training data. Public discussion of this subject frequently centers on massive corpora of text and images that have been scraped from the web,[1] often without consent from the creators of these works.[2] The use of this data is sometimes justified on the basis of the U.S. fair use exception,[3] which, broadly speaking, allows the use of copyrighted materials for secondary purposes that further the aims of copyright and benefit society when this does not disincentivize creative expression.[4] The fair use analysis depends in part on how close the original and secondary use of a work are, whether the secondary use affects the market for the original, and the degree to which the secondary use transforms the original.[5]

---

[1] See e.g., Sheera Frenkel & Stuart A. Thompson, 'Not for Machines to Harvest': Data Revolts Break Out Against A.I., N.Y. Times (July 15, 2023), https://www.nytimes.com/2023/07/15/technology/artificial-intelligence-models-chat-data.html; Elvira Pollina, Italy's privacy regulator looks into online data gathering to train AI, Reuters (Nov. 22, 2023), https://www.reuters.com/world/europe/italys-privacy-regulator-looks-into-online-data-gathering-train-ai-2023-11-22/; Mike Isaac, Reddit Wants to Get Paid for Helping to Teach Big A.I. Systems, N.Y. Times (Apr. 18, 2023), https://www.nytimes.com/2023/04/18/technology/reddit-ai-openai-google.html; Melissa Heikkilä, This new data poisoning tool lets artists fight back against generative AI, MIT Tech. Rev. (Oct. 23, 2023), https://www.technologyreview.com/2023/10/23/1082189/data-poisoning-artists-fight-generative-ai/.

[2] See e.g., Andersen v. Stability AI Ltd., 23-cv-00201, (N.D. Cal. 2023) (where plaintiffs allege that Stability "downloaded … copies of billions of copyrighted images without permission to create Stable Diffusion"); Tremblay v. OpenAI, Inc. 3:23-cv-03223, (N.D. Cal. 2023) Complaint ("OpenAI made copies of Plaintiffs' books during the training process of the OpenAI Language Models without Plaintiffs' permission").

[3] 17 U.S. Code § 107

[4] See Pierre N. Leval, Toward a Fair Use Standard, 103 Harv. L. Rev. 1105, 1110 (1990); Mark A. Lemley & Bryan Casey, Fair Learning, 99 Tex. L. Rev. 743, 760-61 (2020).

[5] See Leval supra note 4 at 1110-25.

Comparatively little attention has been paid to data created specifically for training AI models. This is despite the fact that these specifically curated datasets have been responsible for many of the recent breakthroughs in generative AI. In September 2023, we launched the Data Provenance Initiative, a massive audit of 1,800 curated datasets that traces data back to its original creators and compiles metadata on licensing, dataset characteristics, and dataset creators.[6] Building on this work, we now focus on the legal status of data that is created specifically for AI training. We will discuss how this curated data has enabled recent AI breakthroughs, special considerations for the fair use of curated data, and the interplay between data provenance, copyright protection for data, and responsible AI. From the outset, we acknowledge that our article focuses on U.S. law. We believe this to be justified as many leading generative AI companies are domiciled in the U.S.[7] Thus it is likely that alleged copyright infringement would take place in the U.S. and that the use of training data would be challenged under U.S. copyright law.

2. Curated data has been behind the recent impressive performance of generative AI

AI can learn a great deal from training on massive unstructured data corpora including grammar, lexical semantics, and factual knowledge about the world.[8] Training on unstructured data is generally referred to as pretraining, whereas subsequent training on curated data is known as "finetuning". One of the breakthroughs that gave rise to the latest generation of widely adopted language models like ChatGPT, is the use of highly specialized finetuning datasets created to elicit a wide set of capabilities and enhance the AI's usefulness, helpfulness, and human agreeability in its responses.[9] These specialized datasets include several categories such as domain-specific finetuning, instruction finetuning, and alignment. Domain-specific finetuning equips models with the acumen to generate outputs pertinent to niche areas like law and finance. It not only imparts domain-

[6] Shayne Longpre et al., The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI (2023) [hereinafter Data Provenance], available at https://www.dataprovenance.org/paper.pdf.

[7] Including OpenAI, Anthropic, Cohere (which is also headquartered in Canada), Stability AI (which is also headquartered in the UK).

[8] See Jacob Devlin et al., Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding (2018) (unpublished manuscript), available at https://arxiv.org/pdf/1810.04805.pdf; Alec Radford et al., Language Models are Unsupervised Multitask Learners (2019) (unpublished manuscript), available at https://arxiv.org/pdf/1810.04805.pdf; Fabio Petroni et al., Language Models as Knowledge Bases? (2019) (unpublished manuscript), available at https://arxiv.org/pdf/1909.01066.pdf.

[9] See Long Ouyang et al., Training Language Models to Follow Instructions with Human Feedback, 35 Advances in Neural Info. Processing Sys. 27730 (2022); Victor Sanh et al., Multitask Prompted Training Enables Zero-Shot Task Generalization (2021) (unpublished manuscript), available athttps://arxiv.org/abs/2110.08207; Jason Wei et al., Finetuned Language Models are Zero-Shot Learners, (2021) (unpublished manuscript), available at https://arxiv.org/abs/2109.01652; Hyung Won Chung et al., Scaling Instruction-Finetuned Language Models, (2022) (unpublished manuscript), available at https://arxiv.org/abs/2210.11416.

specific knowledge but also ensures that the outputs are appropriately formatted and contextualized. For instance, a legal AI model would not only parse legal concepts but also communicate them in a manner appropriate to a legal context. Instruction finetuning combines myriad domain-specific finetuning contexts, differentiating them by an "instruction".[10] The goal is to diversify model abilities, enabling responses to a wider array of potential user queries in a useful way. This is particularly relevant for environments like chat interfaces, where models must respond accurately and relevantly to unconstrained user queries.[11] This type of finetuning is crucial for user-centric applications, ensuring the AI's responses are not just correct, but also contextually appropriate. Alignment encompasses methodologies like constitutional AI (RLAIF)[12], direct policy optimization (DPO)[13], and reinforcement learning from human feedback (RLHF)[14]. These approaches are dedicated to aligning AI responses with human preferences, usually to ensure that they reflect an unbiased, polite, and professional tone in practice. This is especially significant in applications where AI interacts directly with users, necessitating outputs that are more cognizant of implicit user expectations and align with societal norms.

An illustrative example of the impact of this curated data is ChatGPT's response to the prompt "Teach an elementary school student about the doctrine of Champerty" (see Appendix). The foundational knowledge about the doctrine originates from the pretraining phase and potentially from domain-specific finetuning with legal data. Meanwhile, the response's structure as a chat-friendly explanation appropriate for a young learner derives from instruction finetuning and alignment tuning.

AI development typically uses much less curated data than pretraining data[15], but the impact of this data is disproportionately large. Each piece of curated data can play a crucial role in shaping the AI's interactive capabilities and our associated perception of its usability. This is in stark contrast to the pretraining phase, where the

[10] See Yizhong Wang et al., Self-Instruct: Aligning Language Model with Self Generated Instructions, (2022) (unpublished manuscript), available at https://arxiv.org/abs/2212.10560 [hereinafter Self-Instruct]; Yizhong Wang et al., Super-naturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks, (2022) (unpublished manuscript), available at https://arxiv.org/abs/2204.07705; Shayne Longpre et al., The flan Collection: Designing Data and Methods for Effective Instruction Tuning, (2023) (unpublished manuscript), available at https://arxiv.org/abs/2301.13688.

[11] See Mina Lee et al., Evaluating Human-Language Model Interaction, (2022) (unpublished manuscript), available at https://arxiv.org/abs/2212.09746.

[12] See Yuntao Bai et al., Constitutional AI: Harmlessness from AI Feedback, (2022) (unpublished manuscript), available at https://arxiv.org/abs/2212.08073.

[13] See Rafael Rafailov et al., Direct Preference Optimization: Your Language Model is Secretly a Reward Model, (2023) (unpublished manuscript), available at https://arxiv.org/abs/2305.18290.

[14] See Ouyang et al. supra note 9.

[15] See e.g., Chunting Zhou et al., Lima: Less is More for Alignment, (2023) (unpublished manuscript), available at https://arxiv.org/abs/2305.11206.

omission of specific data points is less likely to have a noticeable impact on the model's overall efficacy.[16]

Although it has received less attention from legal scholars, curated data plays a critical role in AI systems. The fact that it is created for the sole purpose of AI training, and the way that it affects important aspects of AI systems, affect the application of the fair use doctrine.

3. Fair Use in a Nutshell

The fair use doctrine originated to give courts flexibility in applying copyright law when rigid applications "would stifle the very creativity which that law is designed to foster".[17] The contemporary application of the doctrine considers four factors which are: (1) the purpose and character of the use; (2) the nature of the copyrighted work; (3) the amount and substantiality of the portion copied in relation to the whole; and (4) the effect of the use upon the potential market for or value of the copyrighted work.[18] These factors are given varying weight in the analysis depending on the context.[19]

In a recent submission to the U.S. Copyright Office,[20] we detail how the application of fair use differs between pretraining and curated datasets. We summarize the key considerations below before exploring some additional considerations and complexities.

4. Fair use for pretraining data

Pretraining data consist of content scraped from the web or our sources with minimal additional modification or annotation. While the creators of these datasets likely have a "thin" copyright in the resulting compilation,[21] the key copyright issues revolve around the rights in the underlying works contained in these datasets. Infringement of these rights can occur in three stages of the AI deployment process: first, when the data is initially scraped and *copied* into a dataset, second, when the

[16] See Shayne Longpre et al., A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity, (2023) (unpublished manuscript), available at https://arxiv.org/abs/2305.13169.

[17] Iowa State University v. American Broadcasting, 621 F.2d 57, 60 (2d Cir. 1980)

[18] 17 U.S.C. § 107.

[19] See Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith, 143 S. Ct. 1258, 1274 (2023) ("The Copyright Act's fair use provision . . . 'set[s] forth general principles, the application of which requires judicial balancing, depending upon relevant circumstances.'") (quoting Google LLC v. Oracle Am., Inc., 141 S. Ct. 1183, 1197 (2021)).

[20] U.S. Copyright Office, Comment from Researchers Associated with the Data Provenance Initiative (Nov. 1, 2023), available at https://www.regulations.gov/comment/COLC-2023-0006-9063 [hereinafter Copyright Comment]. Note that in this comment we used "unsupervised data" to refer to pretraining data and "supervised data" to refer to curated data.

[21] See e.g., Feist Publ'ns, Inc. v. Rural Tel. Serv. Co., 499 U.S. 340, 349 (1991) (noting that the copyright in a factual compilation is "thin").

data is used to train a model, and third, when the model is used to generate outputs. Here we will focus on the second issue since many AI developers (especially those at start-ups and smaller companies) rely on other researchers to perform the initial scraping[22] while copyright infringement by AI outputs likely requires a case-by-case analysis.[23]

Though it is a case-specific doctrine, it seems that fair use generally supports the use of pretraining datasets for AI training so long as there is a significant transformation of the underlying works into model weights, limited retention of training data by the model, an aim to extract general insights from the data, and a negligible impact on the original works' market success.[24] Fair use is generally favored when the secondary use of a copyrighted work significantly diverges in purpose from the original.[25] By the nature of unsupervised datasets as compilations of content scraped from sources like blogs, news articles, discussion forums, and online encyclopedias, the original purpose of the works contained in these datasets is quite distant from their use to train AI models. Moreover, because these works were not created to train AI, using them in this way has a small impact on the market for the works. However, it is important to underscore that a trained generative AI model may be used to create outputs that compete with specific works in the training data and this may infringe on the rights of specific creators. Another important consideration relates to the amount and type of content being copied. Pretraining data is intended to give AI models access to generalizable information[26] and this type of factual information is not generally copyrightable.[27] In practice, however, it appears that models may retain significant excerpts of pretraining data,[28] and limiting this type of "recollection" remains an active area of research.[29]

Overall, considering the distinct purposes and markets between the underlying content of pretraining data and its use for machine learning, we expect that courts

[22] For example, the Common Crawl datasets.

[23] But see Peter Henderson et al., Foundation Models and Fair Use, at 1 (2023) (unpublished manuscript) ("If the model produces output that is similar to copyrighted data, particularly in scenarios that affect the market of that data, fair use may no longer apply"), available at https://arxiv.org/pdf/2303.15715.pdf

[24] See Copyright Comment, supra note 20 at 8-11.

[25] See Google LLC v. Oracle Am., Inc., 141 S. Ct. 1183, 1205 (2021) ("The 'substantiality' factor will generally weigh in favor of fair use where, as here, the amount of copying was tethered to a valid, and transformative, purpose.").

[26] Shayne Longpre et al. supra note 16 at 1.

[27] See, e.g., Compendium of U.S. Copyright Practices (Third) § 313.3(A).

[28] See e.g., Peter Henderson et al. supra note 23 at 7; Nicholas Carlini et al., Quantifying Memorization Across Neural Language Models, (2022) (unpublished manuscript), available athttps://arxiv.org/abs/2202.07646.

[29] See e.g., Jooyoung Lee et al., Do Language Models Plagiarize?, in Proceedings of the ACM Web Conference 2023, 3637 (2023).

are likely to generally deem such usage as fair use although this assessment will depend on the details of the generative AI system and training procedure.

5. Fair use for curated data

In contrast to pretraining training data, curated datasets, like those used for finetuning and alignment, are specifically composed and annotated for machine learning purposes. As a result, we argue that the unauthorized use of this data to train generative AI models is less likely to be fair use.[30]

Curated datasets contain expressive content specifically created to instruct generative AI, and so using these for their intended purpose is unlikely to be covered by fair use.[31] In light of the vibrant market for AI training data, the unpermitted use of this training data undermines the commercial potential for training data. We find that 23% of supervised training datasets are published under research or non-commercial licenses,[32] and it is possible that these dataset creators would be willing to permit commercial use of their data in exchange for remuneration. The nature of what a model learns from curated data is also different than pretraining data. Rather than learning largely uncopyrightable information about language and semantics, curated data is often intended to elicit specific types of outputs and thus the models learn to emulate more expressive elements.

Given the fact that curated datasets contain content specifically created for training AI models, along with the market for training data and the different purpose served by these datasets, we expect courts to be less likely to deem the unpermitted usage of these datasets as fair use. However, this determination will likely hinge on a number of additional factors.

6. Not so simple: The interplay between data provenance and fair use

The reality of AI training data is more complex than a simple pretraining-curated data dichotomy. Most datasets do not fit neatly into one category but are the result of multiple steps of scraping, annotation, and compilation by multiple entities and processes. The details of this process are relevant to the fair use analysis and to broader issues around responsible AI development.

Some of these complexities stem from the training data supply chain and the fact that multiple entities contribute to the creation of training data in different ways. In addition, the copyright analysis is complicated by the fact that there are two different types of ownership at stake: ownership in the dataset and ownership of the underlying data. The ownership in the underlying data is especially complicated

[30] See Copyright Comment, supra note 20 at 12-13.
[31] See Warhol, 143 S. Ct. at 1279-82.
[32] See Data Provenance, supra note 6 at 8.

when this data is generated by a model which was itself trained on a range of data. Understanding the provenance of a dataset—where the data originated from, how it was collected and modified, and by whom—is thus critical to assessing how this data may be used.

To illustrate these nuances, consider the ExpertQA[33] dataset that was constructed by asking experts across 32 different fields to formulate questions about their domain. Various Large Language Models (LLMs) were then used to respond to these questions with answers that included references to sources. Finally, experts were asked to rate the quality of responses. What protectable copyright interests do the researchers behind ExpertQA have? Does the use of an LLM for dataset annotations affect these interests? What about the authors of the sources cited in the LLM answers?

In our view, dataset provenance is critical to assessing how data may be used. Based on insights from the Data Provenance Initiative, we highlight three particularly important questions and underscore several areas for future research.

Who created the dataset?

AI developers rarely create all the data they need to train their models, instead they rely on third-party dataset creators, who themselves use data obtained from others to prepare datasets. Understanding the generative-AI supply chain is thus key to answering questions about copyright infringement.[34]

Dataset creators likely have an interest in their dataset as a composition and any expressive content that was added to the dataset. In the case of pretraining data, this is usually just a "thin" copyright interest in the composition of the data and so infringement would require substantial verbatim copying.[35] It remains an open question whether the transformation that occurs when pretraining data is used to train an AI model qualifies as a sufficiently significant to avoid infringement. If so, it may be the case that the licenses associated with pretraining datasets are largely unenforceable against model developers. By contrast, for supervised datasets, dataset creators go to great lengths to create additional content, in the form of annotations and labels, that increase the utility of the dataset for training purposes. As we outline above, it seems likely that this gives them a stronger interest in the resulting dataset.

[33] Chaitanya Malaviya et al., ExpertQA: Expert-Curated Questions and Attributed Answers, (2023) (unpublished manuscript), available at https://arxiv.org/abs/2309.07852.
[34] Katherine Lee et al., Talkin' 'Bout AI Generation: Copyright and the Generative AI Supply Chain(unpublished), available at https://ssrn.com/abstract=4523551.
[35] Experian Info. Solutions v. Nationwide Mktg., 893 F.3d 1176, 1187 (9th Cir. 2018) (finding compilations of factual credit data involved at least minimal creativity, but afforded "thin protection" requiring "substantial verbatim copying").

The characteristics of dataset creators may be relevant to the fair use analysis: we find that 69% of datasets are created by academic institutions while 11% are created by industrial research labs and 11% are created by companies.[36] It seems likely that the unpermitted use of datasets created by for-profit entities would have a particularly significant economic impact. From a responsible AI perspective, it is especially important that commercial AI developers release their training datasets. Enforcing licenses that limit the commercial usage of training data could help incentivize this type of transparency.

Was third-party data used as the basis for the dataset?

Despite the fact that curated datasets contain annotations and labels that were carefully crafted for AI training, they often rely on preexisting data as a starting point. The underlying data is typically sourced from copyrighted content created for purposes other than AI training. For example, a question-answering dataset might start by taking existing articles from Wikipedia and then ask experts to formulate questions that are answered by the article.[37]

If copyrighted data was used as the basis for a dataset, this raises new copyright questions: Is the initial sourcing and processing of copyrighted data to construct curated datasets fair use? At its face, this question appears analogous to the use of pretraining data to train AI models. However, there are several notable distinctions between using copyrighted data for pretraining models versus creating datasets. First, the amount and diversity of preexisting material used in curated datasets is often less than in pretraining data and thus the contribution of each preexisting work is greater. For example, the popular Stanford Question Answering Dataset (SQuAD) dataset contains over 100,000 training examples generated on the basis of just 536 Wikipedia articles.[38] Second, the degree of transformation involved in model training appears greater than in dataset creation since a dataset will often contain verbatim excerpts of the underlying data while a trained model has transformed the training data into model weights. While much focus has centered on utilizing copyrighted data for model training, resolving the questions surrounding constructing datasets in the first place may prove equally consequential.

[36] These figures were obtained via the Data Provenance Explorer, available at https://dataprovenance.org/. Other dataset creators include non-profit research groups and government agencies.

[37] See Pranav Rajpurkar et al., SQuAD: 100,000+ Questions for Machine Comprehension of Text, (2016) (unpublished manuscript), available at https://arxiv.org/abs/1606.05250.

[38] Id.

Were LLMs involved in the creation of the training data?

The practice of utilizing large language models (LLMs) to generate content for AI training is growing; based on our audit, 12% of curated datasets rely on LLMs.[39] There are several reasons why dataset creators might use an LLM to provide annotations. The most obvious reason is that LLMs can be more efficient and affordable than human annotators. LLM outputs from large complex models are also often used to train smaller models, by creating the appropriate training data for a pre-specified purpose.[40] Finally, LLMs can be used to augment human experts by training them to emulate their responses. For example, one application of the ExpertQA dataset is to train a model to generate the expert's ratings of the various LLM outputs and to subsequently use that model to evaluate future outputs to tune them to align with the expert's feedback.

The emerging trend of using LLMs to generate training data prompts raises several important copyright questions. First, the legal ownership of AI-generated content remains ambiguous. While LLM outputs may not be protectable, the process of distilling models into a dataset, sometimes referred to as model exfiltration of model theft,[41] might constitute the extraction of protectable intellectual property. Second, many LLM services' terms forbid using their models to develop competing, third-party systems.[42] While these terms would likely be enforceable against model developers who use LLMs directly to create training data, it is less clear to what degree they would apply when AI developers use data obtained from a third party data creator.[43] Finally, research indicates that training generative models on AI outputs may degrade their performance.[44] As such a copyright regime that requires dataset creators to disclose the use of LLMs for annotation could benefit the AI community more broadly.

As dataset creators increasingly rely on the productive capacity of LLMs for content creation, these issues will gain urgency, and answering the open copyright questions related to model outputs and distilled derivatives will help balance the interests of model creators, dataset curators, and end users. A copyright regime that is grounded in an understanding of the AI development process can help to

[39] See Data Provenance, supra note 6 at 17; see also Steven Y. Feng et al., A Survey of Data Augmentation Approaches for NLP, (2021) (unpublished manuscript), available at https://arxiv.org/abs/2105.03075.
[40] See e.g., Self-Instruct supra note 10.
[41] ee Eduardo Vela, Jan Keller & Ryan Rinaldi, Google's Reward Criteria for Reporting Bugs in AI Products, Google Security Blog (Oct. 26, 2023) available at https://security.googleblog.com/2023/10/googles-reward-criteria-for-reporting.html.
[42] See OpenAI Terms of Use (last accessed Dec. 10, 2023), https://openai.com/policies/terms-of-use.
[43] Data Provenance, supra note 6, at 17.
[44] See e.g., Ilia Shumailov et al., The Curse of Recursion: Training on Generated Data Makes Models Forget, (2023) (unpublished manuscript), available at https://arxiv.org/abs/2305.17493.

incentivize transparency by providing protections for datasets that are shared openly.

7. Looking forward: How dataset provenance can contribute to responsible AI practices

As highlighted by the discussion above, data provenance is important to the fair use analysis, but it is also imperative to responsible AI practices more broadly. AI is fundamentally the product of its training data, and thus understanding the content and limitations of training data is a prerequisite for assessing the limitations of AI systems. By creating a framework where dataset licenses, especially those for curated datasets, are enforceable, regulators can contribute to a where dataset creators are protected when they share their data and model developers must keep track of their data sources.

We find that a majority of curated datasets are created by research entities and shared openly. This openness contributes to transparent and responsible AI practices. Copyright law has an important role to play in incentivizing transparency by giving dataset creators the ability to limit how their work is used via enforceable licenses. Not only can licenses promote transparency, but they can also prevent data from being used inappropriately: dataset creators are well-positioned to evaluate the limitations of their datasets and so their opinions on how their data should be used should be given deference. Licenses represent a convenient way for creators to communicate these limitations and ensure that their work is not misused. While licensing pretraining data has been criticized as unrealistic due to the sheer number of authors whose work is included in pretraining corpora,[45] the relatively small number of supervised datasets makes it much more feasible for developers to obtain licenses from dataset creators.

While the questions we raise are primarily motivated by copyright issues, they also have implications for other areas of AI regulation, including competition and antitrust. Today, much training data is published publicly, a trend likely driven by academic incentives. The enforceability of data licenses could reinforce these incentives and help lower-resourced players train AI models and compete with larger entities. Competition in this space is not limited to AI developers competing with each other, but also extends to AI models competing with the original creators of training data. If preexisting data can be used freely for AI training, this may make it harder for creators to compete with AI models.[46] Finally, the rules surrounding use of

[45] See e.g., Benjamin L.W. Sobel, Artificial Intelligence's Fair Use Crisis, 41 COLUM. J.L. & ARTS 45, 32-33 (2017), available at https://doi.org/10.7916/jla.v41i1.2036.

[46] See e.g., The New York Times Company v. Microsoft Corporation, 1:23-cv-11195, (S.D.N.Y. 2023) Complaint at 49 ("Defendants' generative AI products directly and unfairly compete with Times content and usurp commercial opportunities from The Times.").

model-generated data as training data have significant implications on competition between different model developers and the ability to prevent others from creating distilled derivatives would likely reduce competition among AI developers. More generally, an effective AI competition regime, much like an effective AI copyright regime, must be grounded in an understanding of the AI development process. We leave a full analysis of the interest relationship between training data provenance and competition as an area for future research.

8. Conclusion

We are not the first to point out that legal discourse around AI would benefit from a clear understanding of technical nuances.[47] Copyright law has the potential to incentivize the responsible creation of AI models and training data by fostering transparency and attribution. In our view, enforceable licenses for curated data are an important step towards transparency as they require model developers to determine who created the data they use while protecting data creators who openly share their work. A robust licensing regime thus promotes transparency and incentivizes open science around AI training data. Ultimately, AI systems are predicated on their training data and thus assessing the limitations of AI systems requires understanding the limitations of the training data. By fostering transparency around training data, copyright law can empower efforts to audit AI systems and contribute to safer and more reliable AI research.

Robert Mahari
JD-PhD at Harvard Law School & MIT.

Shayne Longpre
PhD Candidate at the MIT Media Lab.

***

Citation: Robert Mahari and Shayne Longpre, *Discit ergo est: Training Data Provenance And Fair Use*, Dynamics of Generative AI (ed. Thibault Schrepel & Volker Stocker), Network Law Review, Winter 2024.

Acknowledgments

We are immensely grateful to the many contributors involved in the Data Provenance Initiative (www.dataprovenance.org) who made this work possible. We thank Lisette Donewald, Alan Polozov, Ari Lipsitz from the BU/MIT Student Innovation Law Clinic for their legal research and analysi on the fair use analysis. We are grateful to Laura Zhang and Tobin South for their incisive feedback.

[47] See e.g., Katherine Lee et al., supra note 34.

Appendix: ChatGPT’s response to the prompt “Teach an elementary school student about the doctrine of Champerty”



You
Teach an elementary school student about the doctrine of Champerty



ChatGPT
Certainly, Champerty is a legal term that might be a bit complex for an elementary school student, but I'll simplify it as best as I can.

Champerty is an old legal concept that basically means when someone helps another person with a lawsuit in exchange for a share of whatever money or benefits they might win from the case. It's like if you and your friend were playing a game, and you said, "I'll help you win, but in return, I want half of the prize you get if you win."

Now, this idea is usually not allowed in modern law because it could encourage people to get involved in lawsuits just to make money, which isn't always fair. So, most places have rules against champerty to make sure that lawsuits are fair and not influenced by people who are just looking to profit from them.