THE COLUMBIA
SCIENCE & TECHNOLOGY
LAW REVIEW

VOLUME XXVI STLR.ORG FALL 2024

ARTICLE

LOCATION IS ALL YOU NEED: COPYRIGHT EXTRATERRITORIALITY AND WHERE TO TRAIN YOUR AI

Mattias Rättzén[*]

The development of artificial intelligence ("AI") models requires vast quantities of data, which will often include copyrighted materials. The reproduction of copyrighted materials in the course of training AI models will infringe on copyright, unless there are applicable exceptions and limitations exempting such activities. There is so far considerable divergence between jurisdictions, including between the United States, EU, U.K., Japan, Singapore, Australia, India, Israel, and many more countries, in this regard. In the absence of international harmonization, there is therefore a high likelihood that the same type of training activity would be considered copyright infringement in some countries but not in others.

The AI community is not blind to that risk. If copyright law restricts the development and deployment of AI, developers may decide to relocate their operations elsewhere, where the reproduction of training data is clearly not infringing. This Article concludes that there is a loophole in the international copyright system, as it currently stands, that would permit large-scale copying of training data in one country where this activity is not infringing. Once the training is done and the model is complete, developers could then make the model available to customers in other countries, even if the same training activities would have been

[*] LL.M. (Harvard), PGDip (Oxford), LL.M. (Lund). Mattias Rättzén is an intellectual property attorney admitted in New York, England and Wales, and Sweden and the author of several articles on intellectual property law, artificial intelligence, and public and private international law. In addition to working in private practice with intellectual property law in London and Stockholm, he has been a member of the Swedish SCCR delegation at WIPO in Geneva. The author hereby reserves his rights and opt-outs of this Article and its content from being reproduced for text and data mining purposes under Article 4(3) of Directive (EU) 2019/790, and under any other relevant copyright laws that may apply. The author expresses his thanks to the support from the editorial board during the editing of this Article. The views in this Article are the author's own personal views.

infringing if they had occurred there. Because copyright laws are territorial in nature, by default they can only restrict infringing conduct occurring in their respective countries. From that point of view for AI developers, location is indeed all you need.

The EU has become the first to respond to this problem by retroactively extending their text and data mining exception extraterritorially to training activities occurring in non-EU countries, once the completed AI model is placed on the EU market. While such an extraterritorial application benefits rightholders and closes the loophole now present, it makes the situation significantly more complex for developers. If other regulators decide to follow the same path as the EU, which previously happened in the data privacy context, then developers would be facing multiple, conflicting copyright laws targeting the same underlying activity. This could significantly complicate the development process for AI and potentially undermine the AI industry. This Article critically discusses these and related issues, and whether an extraterritorial application of copyright laws is compatible with territoriality norms that are supposed to respect foreign sovereignty. It also explores, in light of these difficulties, whether we should instead shift focus from regulating the inputs (i.e., the data used to train AI models) to regulating the outputs (i.e., the AI-generated content itself). Indeed, to the extent that the transnational data loophole cannot be closed without infringing upon foreign sovereignty, we may need to look at other regulatory means instead.

The Article also suggests that we should consider model training and copyright infringement as a product-by-process problem, which calls for a comparison with how patent law solved similar extraterritoriality issues. Several decades ago, international patent treaties harmonized the extent to which patent laws can be applied extraterritorially to reach imported products derived from foreign manufacturing processes. If regulators wish to extend their copyright laws' extraterritoriality to close the loophole that exists for training activities in the context of AI, and to do so in a way that is aligned with copyright territoriality, there may be a need to similarly revise international copyright treaties. This Article, therefore, urgently calls for a similarly coordinated international effort in copyright law, which balances the interests of rightholders with the technical, regulatory, and economic realities faced by developers. How we resolve these issues could make or break the future of AI. If we cannot find a way to reconcile the interests of rightholders and AI stakeholders, the world may be left with a segregated and fragmented AI landscape, one in which there can only be losers and no winners.

I. INTRODUCTION

Artificial intelligence ("AI"), particularly generative AI, has transformed entire industries within a short period of time, including how we create new content. Tasks historically reserved for human creators can now be automated with remarkable efficiency—generating text, images, music, videos, and code at the click of a button. Data is gold in this fourth industrial revolution.[1] Generative AI models require vast amounts of data to learn and perform their tasks, and it is the relevance, quality, and quantity of that data that determine the model's performance and accuracy, depending on its particular task and use case.[2] The elephant in the (court)room is where AI developers get their data from and how they use it. Data crawling and scraping, which involves collecting information from publicly available sources, is a common initial step in AI development.[3] More often than not, that data will be protected by copyright or related rights. Unsurprisingly, this has spurred an

[1] Mathematician Clive Humby coined the phrase as early as 2006 that "data is the new oil." In 2017, before the recent spike in AI technologies, The Economist declared that the "[t]he world's most valuable resource is no longer oil, but data." *See The world's most valuable resource is no longer oil, but data*, THE ECONOMIST (May 6, 2017), https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data [https://perma.cc/8YQJ-PS6J]. Industry experts agree that data is fundamental for AI and machine learning research. *See*, e.g., Ding Wang, Shantanu Prabhat & Nithya Sambasivan, *Whose AI Dream? In Search of the Aspiration in Data Annotation*, ARXIV 1 (Mar. 21, 2022), https://arxiv.org/abs/2203.10748 [https://perma.cc/8Q4G-GV4T] ("Data is fundamental to AI/ML Models."); LAURA GALINDO, KARINE PERSET & FRANCESCA SHEEKA, OECD, AN OVERVIEW OF NATIONAL AI STRATEGIES AND POLICIES, GOING DIGITAL TOOLKIT NOTE 10-11, https://goingdigital.oecd.org/data/notes/No14_ToolkitNote_AIStrategies.pdf [https://perma.cc/8TU9-VCKV] ("Data access and sharing are key to accelerating AI uptake.").

[2] Arthur Holland Michel, *Recalibrating Assumptions on AI Towards an Evidence-based and Inclusive AI Policy Discourse*, CHATHAM HOUSE 17 (Apr. 2023), https://www.chathamhouse.org/sites/default/files/2023-04/2023-04-05-recalibrating-ai-holland-michel.pdf [https://perma.cc/3JRX-9W7H] (refining the statement that "data is the new oil" and emphasizing that data must be well-curated and free from errors to be of use, as well as closely aligned with the purpose for which the AI system is being developed). *See also* Vivek Iyer et al., *Quality or Quantity? On Data Scale and Diversity in Adapting Large Language Models for Low-Resource Translation*, ARXIV 10 (Aug. 23, 2024), https://www.arxiv.org/abs/2408.12780 [https://perma.cc/2VCT-MZ2V] (analyzing Gemma 2B, Mistral 7B and Llama 3 8B across two language groups, and concluding that "quantity plays the dominant role in downstream performance").

[3] *See infra* Section II.

explosion of copyright lawsuits across the world, alleging that training generative AI models with protected data amounts to copyright infringement.[4]

In December 2023, The New York Times filed a complaint against Microsoft and OpenAI, arguing that ChatGPT was "built by copying and using *millions* of The Times's copyrighted news articles, in-depth investigations, opinion pieces, reviews, how-to guides, and more."[5] Multiple other lawsuits against OpenAI and other AI developers have since been filed. There appears to be no dispute that OpenAI's AI model has been trained on copyrighted materials. Indeed, OpenAI has acknowledged in a submission to the House of Lords in the U.K. that "[b]ecause copyright today covers virtually every sort of human expression—including blogposts, photographs, forum posts, scraps of software code, and government documents—it would be impossible to train today's leading AI models without using copyrighted materials."[6]

Using copyrighted materials when training generative AI models raises difficult questions about ownership, infringement, exceptions and limitations, and calculating damages. These questions have already received significant and well-

[4] Much of the litigation is currently taking place in the United States. *See, e.g.*, Nazemian v. NVIDIA Corp., No. 3:24-cv-01454(N.D. Cal. filed Mar. 8, 2024); Dubus v. NVIDIA Corp., 3:24-cv-02655 (N.D. Cal. filed May 2, 2024); Daily News Corp. v. Microsoft Corp., No. 1:24-cv-03285 (S.D.N.Y. filed Apr. 30, 2024); Doe v. GitHub, Inc., No. 3:22-cv-06823 (N.D. Cal. filed Nov. 3, 2022); Getty Images (US), Inc. v. Stability AI Ltd, No. 1:23-cv-00135 (D. Del. filed Feb. 3, 2023); The Intercept Media, Inc. v. OpenAI, Inc., No. 1:24-cv-01514 (S.D.N.Y. filed Feb. 28, 2024); Raw Story Media, Inc. v. OpenAI, Inc., 1:24-cv-01515 (S.D.N.Y. filed Feb. 28, 2024); Kadrey v. Meta, No. 3:23-cv-03417 (N.D. Cal. filed July 7, 2023); Leovy v. Google, No. 3:23-cv-3440 (N.D. Cal. filed July 11, 2023); Andersen v. Stability AI Ltd., No. 3:23-cv-00201-WHO (N.D. Cal. filed Oct. 30, 2023), Thomson Reuters v. ROSS, No. 1:20-cv-00613 (D. Del. filed May 6, 2020); Concord Music Group, Inc. v. Anthropic PBC, No. 3:23-cv-01092 (M.D. Tenn. filed Oct 18, 2023); Vacker v. ElevenLabs, Inc., 1:24-cv-00987 (D. Del. filed Aug. 29, 2024); UMG Recordings, Inc. v. Suno, Inc., 1:24-cv-11611 (D. Mass. filed June 24, 2024); Huckabee v. Bloomberg,1:23-cv-11195 (S.D.N.Y. filed Dec. 27, 2023); Basbanes v. Microsoft Corp., No. 1:24-cv-00084 (S.D.N.Y. filed Jan. 5, 2024); Sancton v. OpenAI, Inc., 1:23-cv-10211 (S.D.N.Y. filed Nov. 21, 2023). *See also* Getty Images (US) Inc v. Stability AI Ltd [2023] EWHC (Ch) 3090 (U.K.).

[5] Complaint at 2, N.Y. Times v. Microsoft Corp., No. 23-cv-11195 (S.D.N.Y. filed Dec. 27, 2023).

[6] OpenAI, Written evidence to the House of Lords Communications and Digital Select Committee inquiry on large language models, LLM0113, at 4 (Dec. 5, 2023), https://committees.parliament.uk/writtenevidence/126981/pdf [https://perma.cc/L7HT-YVSD].

thought scholarly attention[7] and await answers in ongoing court cases.[8] The discussion in this Article builds on the premise that there is a significant risk that generative AI models, when trained using copyrighted materials, will infringe copyright, and that this will not fall squarely under any existing exceptions or limitations. There is also a significant risk that courts in different countries will come to different conclusions in this regard and that there may be situations where the infringement is excusable depending on the circumstances. Although we continue to wait for pending court cases to be resolved and provide the needed clarity, it is already apparent that there is considerable divergence between copyright laws in different countries regarding text and data mining, including in the United States, EU and the U.K., and among others like Japan, Australia, Singapore, India, and Israel. Whether generative AI models infringe copyright will depend on the facts and the relevant country's laws, presenting a more fragmented picture for developers across the world. Yet as the value of data continues to soar, the need for clarity and consistency becomes increasingly urgent.

If copyright law imposes a barrier to training AI systems on copyrighted materials without authorization from rightholders, then developers may decide to shift their operations to offshore markets where that practice is more clearly permissible. Datasets are stored on servers, and training is performed by people in concert with computers; both can be relocated almost anywhere with little difficulty. We have not seen any development of this sort as of date. But as litigation are brought forth, decisions eventually get laid down, and regulations evolve, it may only be a matter of time until AI developers and their investors decide that their operations can be more safely and reliably managed somewhere else. In a submission before the United States Patent and Trademark Office, OpenAI itself admitted that "[c]opyright barriers to training AI systems would have 'disastrous ramifications' and 'could jeopardize the technology's social value, or drive innovation to a foreign jurisdiction with relaxed copyright constraints.'"[9] This would, of course, be problematic for rightholders whose copyrighted materials are

[7] *See* 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 (2024); Francesca Mazzi & Salvatore Fasciana, *Video Kills the Radio Star: Copyright and the Human Versus Artificial Creativity War*, J. WORLD INTELL. PROP. 1 (2024); Benjamin Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 COLUM. J.L. & ARTS 45 (2017); Katherine Lee et al., *Talkin' 'Bout AI Generation: Copyright and the Generative-AI Supply Chain*, J. COPYRIGHT SOC'Y. U.S.A (forthcoming 2024); Michael D. Murray, *Generative and AI Authored Artworks and Copyright Law*, 45 HASTINGS COMMC'N. & ENT. L.J. 27 (2023); A. Feder Cooper & James Grimmelmann, *The Files are in the Computer: Copyright, Memorization, and Generative AI*, CHI.-KENT L. REV. (forthcoming 2024); Jenny Quang, *Does Training AI Violate Copyright Law?*, 36 BERKELEY. TECH. L.J. 1407 (2021); Frank Pasquale & Haochen Sun, *Consent and Compensation: Resolving Generative AI's Copyright Crisis*, 110 U. VA. L. REV. ONLINE (forthcoming 2024); Amy B. Cyphert, *Generative AI, Plagiarism, and Copyright Infringement in Legal Documents*, 25 MINN. J.L. SCI. & TECH. 49 (2024); Peter Henderson et al., *Foundation Models and Fair Use*, 24 J. MACH. LEARNING RES. 1 (2023).

[8] *See* cases cited *supra* note 4.

[9] OpenAI, LP, Comment Letter on Request for Comments on Intellectual Property Protection for Artificial Intelligence Innovation, 84 Fed. Reg. 58141, at 11 (Jan. 14, 2019) https://www.uspto.gov/sites/default/files/documents/OpenAI_RFC-84-FR-58141.pdf [https://perma.cc/HAP8-JLNB].

being used to train models and represents a loophole in the international copyright system as it stands today. Copyright protection would become worthless if AI developers could relocate their operations to other countries with few or no restrictions on model training, and subsequently sell and market their completed AI models in countries where the same training activity would be an infringement. This loophole has recently come into the spotlight in *Getty Images v. Stability AI* before the English High Court.[10] In that case, Stability AI argued in its defense that there could be no infringement in the U.K. even if Getty's images had been used as data sources for model training without permission, as the processing and hosting services to support the creation of the model had taken place outside the U.K.

Unsurprisingly, regulators have picked up on the fact that this transnational shift in model training could happen. In response, the EU has come to extraterritorially extend the reach of its copyright exception for text and data mining.[11] It remains to be seen whether other countries will follow the footsteps of the EU, but, so far, at least Brazil has indicated its intention to do so.[12] Not long ago, we witnessed a similar trend of extraterritoriality provisions in the context of data privacy, where the EU once again took the first step forward.[13] The concern from regulators is that, if copyright laws are not extended extraterritorially, then models that have been developed based on diverging copyright requirements could end up competing int he same market. If this regulatory trend continues, then it may quickly turn into an unmanageable situation for AI developers who are left to deal with conflicting national copyright laws for the same training activity.

This Article discusses these and related questions for the evolving international copyright regime for AI development, which arise from the fact that diverging copyright exceptions and limitations apply to text and data mining. This Article explores if location indeed is all you need, and to what extent AI developers and providers could escape liability for copyright infringement by offshoring their training activities, therefore taking advantage of the fact that there is no international harmonization on the topic. The Article also considers how copyright extraterritoriality, which so far comes from the EU, could close this transnational data loophole, but critically also whether an extraterritorial application of copyright laws would be compatible with territoriality norms that are supposed to respect foreign sovereignty. Moreover, the Article discusses how copyright

[10] Getty Images (US) Inc v. Stability AI Ltd [2023] EWHC (Ch) 3090.

[11] *See infra* Section VII.A.1. (referring to Recital 106 and Article 53(1)(c) of the Regulation 2024/1689, 2024 O.J. (L 144) 1, 28, 84 (EU) [hereinafter EU AI Act]).

[12] *See infra* Section VII.A.2.

[13] Not just the EU has introduced extraterritorial provisions in Regulation 2016/679, art. 3, 2016 O.J. (L 119) 1, 32-33 [hereinafter GDPR]. Similar provisions can also be found in the data privacy regulations in California (CAL. CIV. CODE § 1798.140 (West 2018)), Singapore (The Personal Data Protection Act 2012, No. 26 of 2012, art. 2(1)), Australia (*Australian Privacy Act 1988* (Cth), s 5B(3)), Brazil (Lei No. 13.709, de 14 de Agosto de 2018, Diário Oficial da União [D.O.U.] de 15.10.2018, ch. 1, art. 3. [hereinafter Brazilian General Data Protection Law 2018]), Canada (A.T. v. Globe24h.com, 2017 F.C. 114 (Can.)), and China (Personal Information Protection Law (promulgated by the 30th meeting of the Standing Comm. Nat'l People's Cong., Aug. 20, 2021, effective Nov. 1, 2021), art. 3), among other countries.

extraterritoriality, if it continues without reservation, could undermine the AI industry, if not dealt with properly, and what better regulatory solutions there could be to address the same issue. It also discusses whether we should, because of these difficulties, consider shifting focus from regulating the inputs (i.e., the data used to train AI models) to regulating the outputs (i.e., the AI-generated content itself). How we resolve these issues could make or break the future of AI. If we cannot find a way to reconcile the interests of rightholders and AI stakeholders, the world may be left with a segregated and fragmented AI landscape, one in which there can only be losers and no winners.

II. A TECHNICAL PRIMER ON TRAINING DATA AND GENERATIVE AI MODELS

Machine learning algorithms used in generative AI models or other deep learning models rely on vast quantities of data to learn patterns and make predictions.[14] The promise of machine learning, and predictive modeling in particular, is that statistical techniques allow for identifying relationships and trends within datasets, which can then be used to inform decisions based on new, unseen information.[15] The success of generative AI models in their ability to make such predictions hinges on the quality, quantity, and representativeness of the data used during training, as well as the algorithms' ability to interpret and act upon the data.

Training generative AI models, particularly large language models ("LLMs") and models used to generate visual or audio content, begins with collection of vast quantities of raw data.[16] Models with different use cases will need different types of data, such as text, images, and videos. For example, OpenAI used publicly accessible datasets such as Common Crawl, WebText2, and Wikipedia for training ChatGPT-3.[17] Common Crawl is a large open repository of web-crawling data, consisting of nearly a trillion words from raw web page data, metadata extracts, and text extracts.[18] However, publicly accessible material does not mean it is free from ownership or copyright. Common Crawl reproduces copyrighted materials in the

[14] Michael R. Douglas, *Large Language Models*, ARXIV 5-7 (Oct. 6, 2023), https://arxiv.org/abs/2307.05782 [https://perma.cc/NBJ7-G426]. *See also* Google LLC, Comment Letter on Artificial Intelligence and Copyright, 88 Fed. Reg. 59942, at 3-4 (Oct. 30, 2023), https://www.regulations.gov/comment/COLC-2023-0006-9003 [https://perma.cc/E2MJ-44S8].

[15] Alfonso Palmer, Rafael Jiménez & Elena Gervilla, *Data Mining: Machine Learning and Statistical Techniques*, *in* KNOWLEDGE-ORIENTED APPLICATIONS IN DATA MINING 373-74 (Kimito Funatsu ed. 2011); MORITZ HARDT & BENJAMIN RECHT, PATTERNS, PREDICTIONS, AND ACTIONS: A STORY ABOUT MACHINE LEARNING 11 (2022) (describing machine learning as the "study of algorithmic prediction" and prediction as "statements about things we don't know for sure yet").

[16] OpenAI, *supra* note 6, at 4; Yiheng Liu et al., *Understanding LLMs: A Comprehensive Overview from Training to Inference*, ARXIV 6 (Jan. 6, 2024), https://arxiv.org/abs/2401.02038 [https://perma.cc/FU29-QXA5] . *See also* Google LLC, *supra* note 14, at 4-5.

[17] Tom B. Brown et al., *Language Models are Few-Shot Learners*, ARXIV 8-9 (July 22, 2020), https://arxiv.org/abs/2005.14165 [https://perma.cc/X5CG-LVNY]. Meta appears to have been using a similar composition of data sources, including Common Crawl, as well as other publicly available content. *See* Susan Zhang et al., *OPT: Open Pre-trained Transformer Language Model*, arXiv, at app. § C.2 (May 2, 2022), https://arxiv.org/abs/2205.01068 [https://perma.cc/Z5W9-PQ29].

[18] Brown et al., *supra* note 17, at 8.

dataset as web pages are archived on a regular basis. WebText2, in contrast to Common Crawl, is an internal dataset created by OpenAI by scraping web pages using its GPTBot with an emphasis on data quality, such as filtered text of web pages linked in upvoted posts on Reddit. Like many other crawlers and scraping tools, the GPTBot supposedly respects robots.txt exclusions.[19] OpenAI supposedly also used books as data sources, although it has declined to comment on what types of books and how they were sourced. In other cases, generative AI developers have been reluctant to disclose the origins of their sourced data. DALL-E, a text-to-image model developed by OpenAI and incorporated into ChatGPT, used more than 400 million image-text pairs from "a variety of publicly available sources on the Internet."[20]

In addition to extracting content from original materials, synthetic data is increasingly used where real-world data is limited, is costly, or raises privacy concerns.[21] Synthetic data refers to artificially generated data using algorithms and simulations that generate data points with similar statistical properties to those of real-world data.[22] Unlike real data, which directly originates from real-world events, synthetic data is created by simulating these events through a purpose-built model.[23] The model can take many different forms and can range from deep learning methods to simpler statistical models or agent-based simulations.[24] Even transformer-based models, which are used in generative AI, can be used to generate synthetic data.[25] Because synthetic data still relies on real-world data being used to

[19] *See Overview of OpenAI Crawlers*, OPENAI, https://platform.openai.com/docs/gptbot [https://perma.cc/3ENX-JST8] (last visited Nov. 12, 2024) ("[A] webmaster can . . . disallow[] GPTbot to indicate that crawled content should not be used for training OpenAI's generative AI foundation models."). Allegations have recently been made that robots.txt are sometimes being ignored, although these allegations have so far not yet been substantiated. *See* Kali Hays, *OpenAI and Anthropic are ignoring an established rule that prevents bots scraping online content*, BUS. INSIDER (June 21, 2004), https://www.businessinsider.com/openai-anthropic-ai-ignore-rule-scraping-web-contect-robotstxt?international [https://perma.cc/X92X-LJYN].

[20] Alec Radford et al., *Learning Transferable Visual Models from Natural Language Supervision*, ARXIV 1 (Feb. 26, 2021), https://arxiv.org/abs/2103.00020 [https://perma.cc/6N7W-LV3W]. Relatedly, a comprehensive review of in total 444 datasets currently available for use, or actually used, by LLMs was recently published. Yang Liu et al., *Datasets for Large Language Models: A Comprehensive Survey*, ARXIV (Feb. 28, 2024), https://arxiv.org/abs/2402.18041 [https://perma.cc/7DJ7-X3X9]. Although the survey does not go into what data is used specifically by various LLM providers, which is not public information, it lists a wide range of possible sources that are currently available, including pre-training corpora, instruction fine-tuning datasets, preference datasets and evaluation datasets. *See id*. at 1.

[21] Shuang Hao et al., *Synthetic Data in AI: Challenges, Applications, and Ethical Implications*, ARXIV 2 (Jan. 3, 2024), https://arxiv.org/abs/2401.01629 [https://perma.cc/C292-QYZB].

[22] James Jordon et al., *Synthetic Data - What, Why and How?*, ARXIV 5 (May 6, 2022), https://arxiv.org/abs/2205.03257 [https://perma.cc/Q523-WBXG].

[23] *Id.*

[24] *Id.*

[25] *See What is Synthetic Data?*, AMAZON WEB SERVICES, https://aws.amazon.com/what-is/synthetic-data/ [https://perma.cc/YN67-6VB5] (last visited Aug. 27, 2024); Dewayne Whitfield, *Using GPT-2 to Create Synthetic Data to Improve the Prediction Performance of NLP Machine Learning Classification Models*, ARXIV (Dec. 31, 2020), https://arxiv.org/abs/2104.10658 [https://

generate larger amounts of data that is statistically similar in some way, it is essential that the real-world data is of high quality. OpenAI and other generative AI developers have reportedly started to rely more on synthetic data based on high quality original data to yield better results.[26] Like any other data, real-world data will often be protected by copyright. Because of that, and because the real-world data will be reproduced in the course of training the model, copyright concerns still arise with respect to synthetic data.

Once the data is collected, the next step is data pre-processing to ensure relevance, consistency, and accuracy.[27] Pre-processing steps include filtering out low-quality , irrelevant, conflicting, biased, or harmful data, such as "toxic" speech.[28] The pre-processing step not only sorts out unwanted and outlier data, but it also consistently formats the data in a manner that is compatible with the specific model.[29] The vast quantities of data make supervised machine learning impractical in this regard, and therefore, unsupervised models are typically employed.[30] Once the data has been pre-processed , another common step in the data lifecycle process is data transformation. Instead of removing or filtering out particular data, this involves manipulating the data itself.[31] Depending on the use case, strategies for data transformation include smoothing by removing noise, assigning attributes, normalization, discretization, or aggregation.[32] Data augmentation techniques are often used to improve the relevance and quality of the original data.[33] For example, inconsistent values found in the original data can be replaced with synthetic ones, a method known as data curation. The same method can be used to fill gaps in the

perma.cc/7DJ7-X3X9]; Aivin V. Solatorio & Olivier Dupriez, *REaLTabFormer: Generating Realistic Relational and Tabular Data using Transformers*, ARXIV (Feb. 4, 2024), https://arxiv.org/abs/2302.02041 [https://perma.cc/7Z33-G9GD]; Xu Guo & Yiqiang Chen, *Generative AI for Synthetic Data Generation: Methods, Challenges and the Future*, ARXIV (Mar. 7 2024), https://arxiv.org/abs/2403.04190 [https://perma.cc/W2LC-7ZDB].

[26] *See, e.g.*, Brown et al., *supra* note 17, at 21; Lakshmi Varanasi, *As most AI execs scramble for more data, Mark Zuckerberg says there's actually something more 'valuable'*, BUS. INSIDER (Apr. 21, 2024), https://www.businessinsider.com/mark-zuckerberg-meta-ai-model-training-synthetic-data-feedback-loops-2024-4.

[27] Humza Naveed et al., *A Comprehensive Overview of Large Language Models*, ARXIV 5 (Apr. 9, 2024), https://arxiv.org/abs/2307.06435 [https://perma.cc/F8LG-P22C]; Martin Kretschmer, Thomas Margoni & Pinar Oruç, *Copyright Law and the Lifecycle of Machine Learning Models*, 55 INT'L REV. INTELL. PROP. & COMPETITION L. 110, 118 (2024).

[28] Naveed et al., *supra* note 27, at 5; Liu et al., *supra* note 16, at 8.

[29] Alvaro A. Fernandes et al., *Data Preparation: A Technological Perspective and Review*, 4 SN COMPUT. SCI. 425, at 425 (2023), https://doi.org/10.1007/s42979-023-01828-8 [https://perma.cc/UQ2J-79MV].

[30] Shayne Longpre et al., *A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity*, ARXIV 3 (Nov. 13, 2023), https://arxiv.org/abs/2305.13169 [https://perma.cc/4RQZ-UJW6].

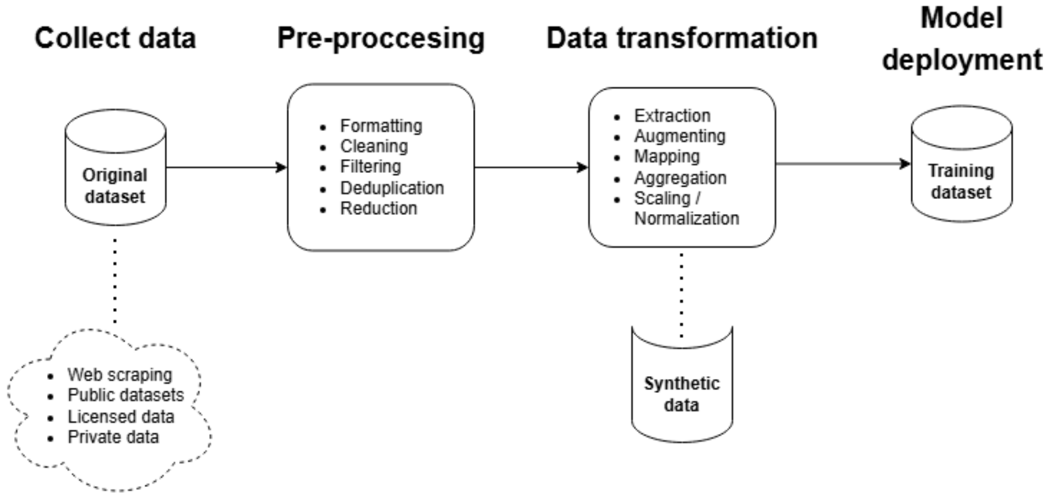
[31] JIAWEI HAN, MICHELIN KAMBER & JIAN PEI, DATA MINING, CONCEPTS AND TECHNIQUES 111-13 (3d ed. 2011).

[32] *Id.* at 112.

[33] Tian Yu Liu & Baharan Mirzasoleiman, *Data-Efficient Augmentation for Training Neural Networks*, ARXIV 1-2 (Oct. 15, 2022), https://arxiv.org/abs/2210.08363 [https://perma.cc/A3FK-F4FW].

original dataset to improve information extraction and is frequently used to remove personal data.[34] The original web pages or data files therefore rarely form part of the dataset. Typically, some level of extraction or processing of the data will have taken place, either at the initial stage of data collection—for example when scraping web pages for relevant content—[35] and/or when pre-processing the collected data.

Figure 1: Common data pipeline for machine learning models



A critical aspect of training generative AI models involves the use of tokens.[36] For text-based models, data is broken down into smaller units called a token, which can be a whole word or a part of word.[37] ChatGPT-3, for example, was trained on more than 300 billion tokens.[38] These tokens are transformed into numerical representations that the model can process, a process known as encoding.[39] How encoded tokens are used more precisely will differ depending on the model architecture and design. Currently, all major LLMs, including ChatGPT, Llama and Mistral 7B, are built on the transformer architecture, illustrated in Figure 2.[40] In transformer-based models, attention heads assign weights to input tokens.[41] Transformers use a self-attention structure to compute the representation of a token sequence, which encodes the position of each token sequence through a process called positional embedding.[42] Positional embedding involves assigning a unique vector to each token sequence.[43] The attention mechanism in the transformer model

[34] Liu et al., *supra* note 16, at 8; Jordon et al., *supra* note 22, at 30-31.

[35] Zhipeng Xu et al., *Cleaner Pretraining Corpus Curation with Neural Web Scraping*, ARXIV 2, (June 15, 2024), https://arxiv.org/abs/2402.14652 [https://perma.cc/G47A-UUY3].

[36] Naveed et al., *supra* note 27, at 4.

[37] *Id.* at 4.

[38] Brown et al., *supra* note 17, at 8.

[39] Liu et al., *supra* note 16, at 3.

[40] *Id.* at 9.

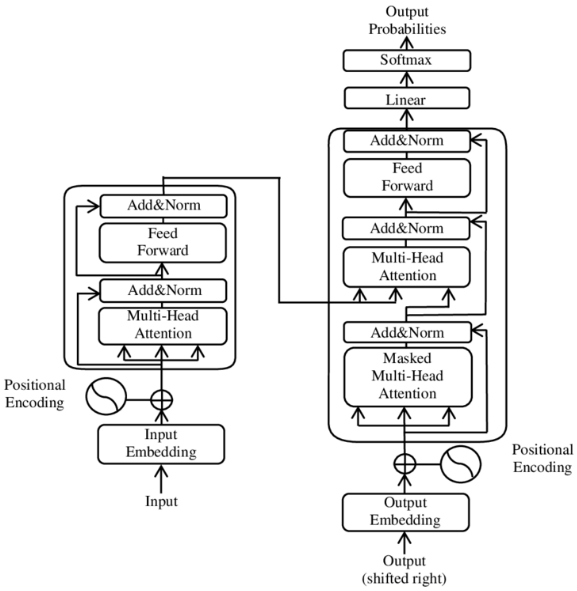
[41] *Id.* at 2; Naveed et al., *supra* note 27, at 4.

[42] Liu et al., *supra* note 16, at 2.

[43] *Id.* at 3.

then uses these embeddings to weigh the relevance of each token in the context of the entire sequence.[44] This allows the model to relate a single token with other tokens and differentiate between different tokens based on their positions in the sequence.[45] It is from this so-called pre-training process that the model learns statistical relationships between tokens.[46] The statistical relationship is encoded into parameters, including subsets known as weights, which determine the relative strength of connections between different tokens.[47] The number of these weights are adjusted during training, both during pre-training and fine-tuning, to tell when the neural network is to be activated.[48] By evaluating the outcome of the output, it is possible to tweak the model to produce more accurate or relevant results.

Figure 2: Transformer model architecture[49]



Through this process, the model will "learn" to produce decoded output, which is statistically representative of the token patterns found in the training data and transformed into parameters. When generative AI models output content based on a prompt, they are therefore predicting what is the next likely word in a given context.[50] The pre-training phase itself is an iterative process where the model ingests more data and adjusts its internal parameters accordingly to minimize error

[44] *Id*. at 2. *See also* Ashish Vaswani et al., *Attention Is All You Need*, ARXIV 2-3 (Feb. 28, 2024), https://arxiv.org/pdf/2402.18041 [https://perma.cc/LS7L-XRJC].

[45] Vaswani et al., *supra* note 44, at 5.

[46] *Id.* at 6-7; Liu et al., *supra* note 16, at 10; Douglas, *supra* note 14, at 19-22.

[47] Liu et al., *supra* note 16, at 9-10.

[48] Yann LeCun, Yoshua Bengio & Geoffrey Hinton, *Deep Learning*, 521 NATURE 436, 436-37 (2015).

[49] Figure obtained from Yuening Jia, *Attention Mechanism in Machine Translation*, J. PHYS.: CONF. SERIES, Nov. 6, 2019, at 1, 3 fig. 1.

[50] Liu et al., *supra* note 16, at 5, 10.

in its predictions.[51] For example, in LLMs, the model predicts the next word in a sequence based on the context provided by the preceding words.[52] Similarly, if a generative image model is trained on a dataset of images, then that model will seek to generate new images that are statistically similar to those in the training set. Simply put, generative AI models function by identifying and assessing statistical patterns from the training data.[53] Of course, this also means that if the training data is incorrect, inconsistent, or of poor quality, then that same shortcoming will be reflected in the generated output. The model parameters adjusted during training are stored and used when deploying the model to generate outputs.[54] The parameters do not store the training data itself but rather the learned patterns and statistical relationships derived from the data.[55] Whether generative AI models "memorize" the learned data, meaning that the data can be reproduced identically or nearly identically, has become a contentious issue in several of the pending litigations.[56] There is evidence that indicates models such as ChatGPT and Stable Diffusion may have memorized copyrighted materials,[57] which raises additional copyright concerns.

One of the drivers behind the most recent breakthrough for LLMs was the substantial increase of parameters for transformer capacity. Previous transformer-based LLMs used only a few billion parameters.[58] What turned out to make a significant difference in text synthesis and downstream natural-language processing tasks when ChatGPT-3 was developed was the substantial increase of transformer capacity from a few billion parameters, to 175 billion parameters, ten times more than any previous non-sparse language model.[59] In the influential research paper that laid out the foundation of ChatGPT-3, it was concluded that "a 175 billion parameter language model . . . shows strong performance on many NLP tasks and benchmarks in the zero-shot, one-shot, and few-shot settings, in some cases nearly matching the performance of state-of-the-art fine-tuned systems, as well as generating high-quality samples and strong qualitative performance at tasks

[51] *Id.* at 3.

[52] Naveed et al., *supra* note 27, at 11; Brown et al., *supra* note 17, at 11.

[53] Ivo Emanuilov & Thomas Margoni, *Forget Me Not: Memorisation in Generative Sequence Models Trained on Open Source Licensed Code*, ZENODO 10 (2024), https://zenodo.org/records/10635479 [https://perma.cc/XB7T-5ZTD].

[54] *Id.* at 5; Liu et al., *supra* note 16, at 10.

[55] Liu et al., *supra* note 16, at 10.

[56] *See, e.g.*, Complaint at 29-30, N.Y. Times Co. v. Microsoft Corp., No. 23-cv-11195 (S.D.N.Y. filed Dec. 27, 2023).

[57] Exhibit J, N.Y. Times Co. v. Microsoft Corp., No. 23-cv-11195 (S.D.N.Y. filed Dec. 27, 2023); Carlini et al., *Extracting Training Data from Diffusion Models*, ARXIV (Jan. 30, 2023), https://arxiv.org/pdf/2301.13188 [https://perma.cc/BM3D-9CZL]. The issue of "memorization" and data encoding is discussed in more detail below. *See infra* Section III.C.

[58] Brown et al., *supra* note 17, at 4 ("[I]n recent years the capacity of transformer language models has increased substantially, from 100 million parameters [RNSS18], to 300 million parameters [DCLT18], to 1.5 billion parameters [RWC+19], to 8 billion parameters [SPP+19], 11 billion parameters [RSR+19], and finally 17 billion parameters [Tur20].").

[59] *Id.* at 5.

defined on-the-fly."[60] This proved the hypothesis that "[s]ince in-context learning involves absorbing many skills and tasks within the parameters of the model, it is plausible that in-context learning abilities might show similarly strong gains *with scale*."[61] The reason for this is still not fully understood technically. However, what is now clear is that larger foundation models with more parameters, which are derived from larger quantities of training data, are able to capture more complex patterns found in the training data, yielding better performance.[62] This has come to be known as the "emergence" phenomenon, which is uniquely present in larger models.[63] Why this phenomenon can only be witnessed for larger models largely remains a technical mystery to this date.[64]

That is not to say that quantity of data is everything. The quality and relevance of the data also matters as much, if not more,[65] particularly where larger foundation models are to be deployed for specific use cases or tasks.[66] Today, more specialized generative AI models are typically developed by either building and training from scratch using large datasets or fine-tuning existing pre-trained models. There are several different methods for fine-tuning pre-trained models, which can use either a domain-specific or mixed approach. Domain-specific fine-tuning uses new, specialized training data to enhance the model's performance in a particular field.[67] This poses the risk, however, of so-called "catastrophic forgetting" or performance degradation, where the model experiences a moderate to severe decline in performance on previously learned tasks.[68] Simply put, the fine-tuned adjustments

[60] *Id.* at 40-41.

[61] *Id.* at 4 (emphasis added).

[62] Liu et al., *supra* note 16, at 1-2; Jared Kaplan et al., *Scaling Laws for Neural Language Models*, ARXIV 2-6 (Jan. 23, 2020), https://arxiv.org/abs/2001.08361 [https://perma.cc/SM3N-4B4L]. This has also been repeatedly explained by Ilya Sutskever, co-founder of and former chief scientist at OpenAI, in interviews. *See, e.g.*, *Fireside Chat with Ilya Sutskever and Jensen Huang*, NVIDIA ON-DEMAND (Mar. 22, 2023), https://www.nvidia.com/en-us/on-demand/session/gtcspring23-s52092/.

[63] Jason Wei et al., *Emergent Abilities of Large Language Models*, ARXIV 1-2 (Oct. 26, 2022), https://arxiv.org/abs/2206.07682 [https://perma.cc/GE9H-BBFW; Hang Chen et al., *Quantifying Emergence in Large Language Models*, ARXIV 1-2 *(*May 21, 2024), https://arxiv.org/abs/2405.12617 [https://perma.cc/F6X8-EPFM].

[64] *Id.* at 7-8.

[65] *See supra* note 2.

[66] Ziche Liu et al., *Take the Essence and Discard the Dross: A Rethinking on Data Selection for Fine-Tuning Large Language Models*, ARXIV 1-2 (June 20, 2024), https://arxiv.org/abs/2406.14115 [https://perma.cc/NP98-HCFP]; Chunting Zhou, *LIMA: Less Is More for Alignment*, ARXIV 1-2 (May 18, 2023), https://arxiv.org/abs/2305.11206 [https://perma.cc/X55Y-79MU]; Hugo Touvron et al., *Llama 2: Open Foundation and Fine-Tuned Chat Models*, ARXIV 9 (July 19, 2023), https://arxiv.org/abs/2307.09288 [https://perma.cc/BD4N-JY8J].

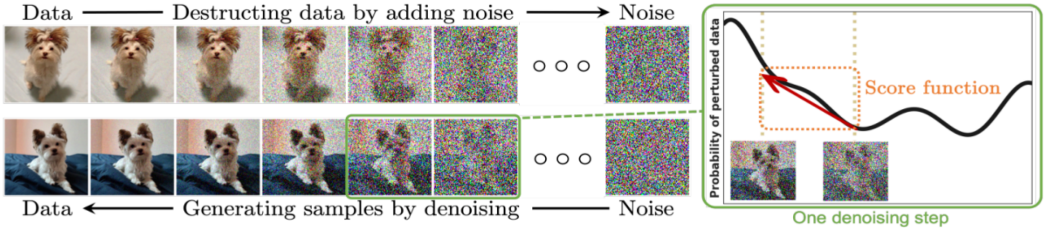
[67] Jiawei Zheng et al., *Fine-tuning Large Language Models for Domain-specific Machine Translation*, ARXIV 2 (Feb. 23, 2024), https://arxiv.org/abs/2402.15061 [https://perma.cc/RML7-KGN4]; Cheonsu Jeong, *Fine-tuning and Utilization Methods of Domain-specific LLMs*, ARXIV 7-8 (Jan. 1, 2024), https://arxiv.org/abs/2401.02981 [https://perma.cc/3RDA-AN7D].

[68] Chengyuan Liu et al., *More Than Catastrophic Forgetting: Integrating General Capabilities For Domain-Specific LLMs*, ARXIV 1-2 (May 28, 2024), https://arxiv.org/abs/2405.17830 [https://perma.cc/VS8R-7XXD].

incorrectly overwrite or disrupt the neural network's representations of previously learned information. Mixed data fine-tuning seeks to prevent that by combining original and new data, which helps the model retain its original knowledge while learning new information.[69] Using a pre-trained model does not always necessitate access to the original training data if only domain-specific fine-tuning is employed, in which case new, typically more specialized data is used instead. However, mixed data fine-tuning will typically require access to the original dataset in order to ensure compatibility and to reformat or improve the training process.[70]

Although transformer-based models like LLMs have become mainstream for use in generative AI, diffusion-based models are also often used. Diffusion-based models are particularly common for image generation and are used in both Midjourney's model and Stability AI's model, Stable Diffusion.[71] Diffusion-based models typically use a convolutional autoencoder network combined with transformer-based text encoders.[72] The autoencoder employs Denoising Diffusion Probabilistic Models ("DDPM") to manipulate latent image vectors by iteratively adding and removing Gaussian noise.[73] Initially, the encoder converts an image into a latent vector, which is then corrupted by progressively increasing noise levels at different timesteps while reducing image resolution and teaching the noise predictor how much noise was added.[74] The original image becomes virtually incomprehensible at the final noise stage, as illustrated in Figure 3.

Figure 3: Encoding and decoding process of diffusion-based models[75]


Data
Destructing data by adding noise
Noise
Probability of perturbed data
Score function
Data
Generating samples by denoising
Noise
One denoising step

[69] Guanting Dong et al., *How Abilities in Large Language Models are Affected by Supervised Fine-tuning Data Composition*, ARXIV 4-8 (Oct. 9, 2023), https://arxiv.org/abs/2310.05492 [https://perma.cc/4U8D-38UG].

[70] Yasmin Moslem, Language Modelling Approaches to Adaptive Machine Translation 29, 38-42 (Ph.D. dissertation, Dublin City University, 2024).

[71] Greg Schoeninger, *ArXiv Dives - Diffusion Transformers*, OXEN (Mar. 12, 2024), https://www.oxen.ai/blog/arxiv-dives-diffusion-transformers [https://perma.cc/A9N3-R3VC].

[72] William Peebles & Saining Xie, *Scalable Diffusion Models with Transformers*, ARXIV 4 (Mar. 2, 2023), https://arxiv.org/abs/2212.09748 [https://perma.cc/ET9L-VJQN].

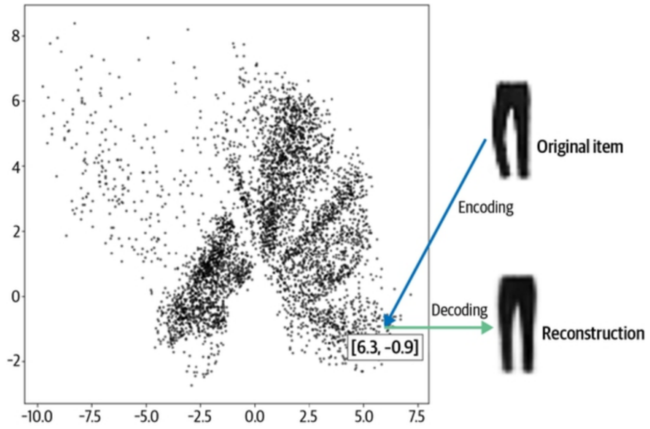
[73] *Id.* at 3-4.

[74] *Id.* at 3; Xiefan Guo et al., *INITNO: Boosting Text-to-Image Diffusion Models via Initial Noise Optimization*, ARXIV 3 (Apr. 6, 2024), https://arxiv.org/abs/2404.04650 [https://perma.cc/98CE-9BF9].

[75] Ling Yang et al., *Diffusion Models: A Comprehensive Survey of Methods and Applications*, ARXIV 6 fig. 2 (June 24, 2024), https://arxiv.org/abs/2209.00796 [https://perma.cc/DX83-S2YQ].

The decoder reverses this process, gradually removing the noise to reconstruct the original image.[76] Score-based generative models will estimate the score function for each denoising step in the reverse process.[77] The score function is defined as a gradient, which is a vector field pointing to different directions with higher likelihood of data and less noise.[78] This way, diffusion-based models are able to compress an original image into nothing more than a numerical representation on the latent space vector, as illustrated in Figure 4.

Figure 4: Latent space vector of an encoded original image[79]



The numerical representation in the latent space vector performs a similar function to that of a parameter or weight in a transformer-based model, which are adjusted during the pre-training process. The so-called hyperparameters are also adjusted during the pre-training process, which include noise schedule, the number of diffusion steps, learning rate, batch size, etc.[80] Text-to-image diffusion-based models use text encoders to prompts to generate latent descriptions as tokens, which are fused with the decoder input, conditioning the image generation on text relevance.[81] Text tokens associate the meaning of text with a corresponding image through attention heads.[82] During sampling, noise vectors seed the decoder, which

[76] Guo et al., *supra* note 74, at 3.

[77] Yang Song et al., *Score-Based Generative Modeling through Stochastic Differential Equations*, arXiv 1-2 (Feb. 10, 2021), https://arxiv.org/abs/2011.13456 [https://perma.cc/R8FB-ZXU7].

[78] *Id.* at 6.

[79] Schoeninger, *supra* note 71.

[80] Peebles & Xie, *supra* note 72, at 4, 6-7; Chen Henry Wu & Fernando De la Torre, *Unifying Diffusion Models' Latent Space, with Applications to Cycle Diffusion and Guidance*, arXiv (Dec. 7, 2022), https://arxiv.org/abs/2210.05559 [https://perma.cc/B3YF-VLR9].

[81] Arman Zarei et al., *Understanding and Mitigating Compositional Issues in Text-to-Image Generative Models*, arXiv 1-2 (June 12, 2024), https://arxiv.org/abs/2406.07844 [https://perma.cc/DU2Q-DAXD].

[82] Kaiyi Huang et al., *T2I-CompBench: A Comprehensive Benchmark for Open-world Compositional Text-to-image Generation*, arXiv 4 (Oct. 30, 2023), https://arxiv.org/pdf/2307.06350 [https://perma.cc/43AH-KDU4]; Arman Zarei et al., *Understanding and Mitigating Compositional Issues in Text-to-Image Generative Models*, arXiv 4-5 (June 12, 2024) https://arxiv.org/pdf/2406.07844 [https://perma.cc/UX76-6B5G].

denoise output at each timestep based on text encoding guidance.[83] When prompted, it produces a random tensor in the latent space, so-called latent noise.[84] During the decoding phase, the diffusion model predicts the noise in the latent space, which it had previously been taught during the encoding phase.[85] Rather than adding or removing noise directly to the original image, it therefore reconstructs the latent representation in the latent space, allowing it to progressively reconstruct original images from noisy latent representations.[86] This decoding process is guided by text encoders, which provide semantic context.

Once the diffusion-based model is trained, the model no longer needs access to the original images that were used for generating new samples. The model will start with random noise and, using the learned parameters, will reverse the diffusion process by predicting the latent noise to generate the image.[87] The reproduction is done in the latent space only. The benefit of doing so is that the memory and compute requirements are substantially reduced.[88] But this also means the original dataset is not required during sampling, as the model relies only on its learned ability to map noise. Although diffusion-based models work differently than transformer-based models in many ways, they similarly require vast quantities of data to learn image and corresponding text representations in the latent space.[89] For example, Stable Diffusion has been trained on the LAION-5B dataset, which consists of 5.85 billion CLIP-filtered image-text pairs.[90] This dataset was created by starting from Common Crawl by extracting and downloading image-text pairs from nearly 3 billion crawled, publicly available web pages.[91] Again, and to state the obvious, the fact that web pages are publicly available does not mean they are free from copyright.

[83] Yang et al., *supra* note 75, at 6-10.

[84] *Id.*

[85] Peebles & Xie, *supra* note 72, at 5; Song et al., *supra* note 77, at 5-7.

[86] Yang et al., *supra* note 75, at 20-21; Peebles & Xie, *supra* note 72, at 5.

[87] Song et al., *supra* note 77, at 7.

[88] Robin Rombach et al., *High-Resolution Image Synthesis with Latent Diffusion Models*, ARXIV 3 (Apr. 13, 2022), https://arxiv.org/pdf/2112.10752 [https://perma.cc/648X-ZNB9].

[89] Zhendong Wang et al., *Patch Diffusion: Faster and More Data-Efficient Training of Diffusion Models*, ARXIV 2 (Oct. 18, 2023), https://arxiv.org/pdf/2304.12526 [https://perma.cc/N979-3ZQS]; Guangkai Xu et al., *Diffusion Models Trained with Large Data Are Transferable Visual Models*, ARXIV 1 (Mar. 10, 2024), https://arxiv.org/pdf/2403.06090v1 [https://perma.cc/NC6R-V37P].

[90] Christoph Schuhmann et al., *LAION-5B: An Open Large-Scale Dataset For Training Next Generation Image-Text Models*, ARXIV 1 (Oct. 16, 2022), https://arxiv.org/pdf/2210.08402 [https://perma.cc/VA8E-YR2F].

[91] *Id.* at 4-5.

III. THE COPYRIGHT INFRINGEMENT DILEMMA FOR MODEL TRAINING AND ENCODING

A. The Digital Right of Reproduction in the Data Supply Chain

Copyright protects its owner against unauthorized copying. Under Article 9(1) of the Berne Convention,[92] authors universally enjoy the exclusive right to reproduce their works, which is triggered whenever a work is "reproduced," in any manner or form.[93] In the United States, owners of copyright have the exclusive right to "reproduce" the copyright work in "copies."[94] Copies are defined in the U.S. Copyright Act as "material objects . . . in which a work is fixed by any method now known or later developed, and from which the work can be perceived, reproduced, or otherwise communicated, either directly or with the aid of a machine or device."[95] Similarly, in the U.K., the owner of the copyright has the exclusive right to "copy" the work in the U.K.[96] The reproduction right is also harmonized at an EU-level, granting authors the exclusive right to authorize or prohibit the "direct or indirect, temporary or permanent reproduction by any means and in any form, in whole or in part."[97]

It is clear that the right of reproduction extends to copies of works that are reproduced in digital form.[98] So long as a work has been fixated digitally,[99] and so

[92] Berne Convention for the Protection of Literary and Artistic Works, Sept. 9, 1886, as revised at Paris, July 24, 1971, and as amended on Sept. 28, 1979, S. Treaty Doc. No. 99-27, 1161 U.N.T.S. 3 [hereinafter Berne Convention].

[93] Specifically, Article 9(1) of the Berne Convention, *supra* note 92, provides that authors of literary and artistic works shall have "the exclusive right of authorizing the reproduction of these works, in any manner or form."

[94] 17 U.S.C. § 106(1).

[95] 17 U.S.C. § 101 (defining "copies").

[96] Copyright, Designs and Patents Act 1988, c. 48, § 16(1)(a) (U.K.).

[97] Council Directive 2001/29, art. 2(a), 2001 O.J. (L 167) 10, 16 (EU) [hereinafter Infosoc Directive].

[98] Although not strictly binding as an integral part of international treaties, WIPO signatory states to the WIPO Copyright Treaty issued an agreement statement at a diplomatic conference on December 20, 1996, that "[t]he reproduction right, as set out in Article 9 of the Berne Convention, and the exceptions permitted thereunder, fully apply in the digital environment, in particular to the use of works in digital form. It is understood that the storage of a protected work in digital form in an electronic medium constitutes a reproduction within the meaning of Article 9 of the Berne Convention." Agreed Statement concerning the WIPO Copyright Treaty, Dec. 20, 1996, WIPO Lex No. TRT/WCT/002, https://www.wipo.int/wipolex/en/text/295456 [https://perma.cc/WGE9-HANC]. *See* Caterina Sganga, *The right of reproduction*, *in* RESEARCH HANDBOOK ON EU COPYRIGHT LAW ch. 6, at 7 (Eleonora Rosati ed. 2021); Lee et al., *supra* note 7, at 67-68. *See also* ZOHAR EFRONI, ACCESS-RIGHT: THE FUTURE OF DIGITAL COPYRIGHT LAW 203-47 (2010) (summarizing the position of a digital reproduction right in international copyright treaties, the United States and the EU). In the U.K., Copyright, Designs and Patents Act 1988, c. 48 § 17(2) further defines infringement by copying as "reproducing the work in any material form. This includes storing the work in any medium by electronic means."

[99] 17 U.S.C. § 101 limits the term "copies" to the "material object . . . in which the work is first fixed." A work becomes "fixed" when it is "sufficiently permanent or stable to permit it to be perceived, reproduced, or otherwise communicated for a period of more than transitory duration."

long as it does not become a mere transient reproduction or is an otherwise exempt activity,[100] an infringing act will have taken place. The right of reproduction is therefore inherently broad and meant to extend to any type of copying, regardless of what medium is used and in what format the work is being copied. There is little doubt that, if copyrighted materials have been downloaded and stored for the purpose of creating datasets, then digital copies will have been created in the process. Because those copies are not merely transient or temporary, but fixed in a material object and form, an act of reproduction will have occurred.[101] The exception for temporary acts of reproduction in the Infosoc Directive in the EU has been narrowly interpreted by the European Court of Justice, which makes it limited to copies that have no independent economic value and that are deleted automatically from the computer memory.[102] This permits, for example, cache copies made during internet browsing,[103] but that is very different from

Id. *See also* Cartoon Network LP v. CSC Holdings, Inc., 536 F.3d 121, 127 (2d Cir. 2008) (holding that "copies" must therefore remain embodied "for a period of more than transitory duration" to infringe). Fixation is not a strict requirement in U.K. or EU copyright law; however, it becomes indirectly relevant through exceptions and limitations that apply to temporary or transient copies of works. *See* EFRONI, *supra* note 98, at 242-43.

[100] Infosoc Directive, *supra* note 97, at art. 5(1) (EU) (excluding temporary acts of reproduction from infringing, which are transient or incidental and an integral and essential part of a technological process, and whose sole purpose is to enable a transmission in a network between third parties by an intermediary, or a lawful use, and where the work has no independent economic significance); CDPA § 28A (U.K.) (excluding the making of temporary copies from infringement on the same basis as the Infosoc Directive, *supra* note 97). In the United States, 17 U.S.C. § 117(a)(1) exempts owners of a copy of a computer program from infringing copyright for copies of computer programs that are generated as an "essential step in the utilization of the computer program in conjunction with a machine and that . . . is used in no other manner." This is, however, strictly limited to owners of "computer programs." *See* Dena Chen et al., *Providing an Incidental Copies Exemption for Service Providers and End-Users*, SAMUELSON L., TECH. & PUB. POL'Y CLINIC, UC BERKELEY SCH. OF L. 8-9 (Mar. 31, 2011), https://www.law.berkeley.edu/wp-content/uploads/2015/04/craincidentalcopies.pdf [https://perma.cc/J8WQ-D8LJ].

[101] Lee et al., *supra* note 7, at 68 (stating that "all [models and generations] trigger the reproduction right when they are created, because they are stored in material objects"). *See also* MAI Sys. Corp. v. Peak Computer, Inc., 991 F.2d 511, 519 (9th Cir. 1993) (finding that "the loading of software into a computer constitutes the creation of a copy under the [U.S.] Copyright Act"); Capitol Records, LLC v. ReDigi Inc., 910 F.3d 649, 657 (2d Cir. 2018) (finding that "[t]he fixing of the digital file [in a] server, as well as in the new purchaser's device, creates a new phonorecord, which is a reproduction")

[102] In *Infopaq I*, the European Court of Justice clarified that temporary and transient acts of reproduction must be "intended to enable the completion of a technological process of which it forms an integral and essential part." *See* Case C-5/08, Infopaq Int'l A/S v Danske Dagblades Forening (*Infopaq I*), 2009 E.C.R. I-6624, ¶ 61. This was narrowed down even further in *Infopaq II*, where the Court explained that temporary acts of reproduction "must pursue a sole purpose, namely the lawful use of a protected work or a protected subject-matter," must not "have an independent economic significance," and must be "automated so that it deletes that act automatically, without human intervention, once its function of enabling the completion of such a process has come to an end." *See* Case C-302/10, Infopaq Int'l A/S v Danske Dagblades Forening (*Infopaq II*), ECLI:EU:C:2012:16, ¶¶ 46, 54, 58 (Jan. 17, 2012). *See also Infopaq I*, at ¶ 64.

[103] *Infopaq I*, 2009 E.C.R. ¶ 63 (citing Recital 33 of the Infosoc Directive, *supra* note 97).

downloading and storing copies for the purpose of permanently including them in datasets.

An act of reproduction will amount to copyright infringement, unless only individual fragments have been copied, in such a way that these fragments are no longer identical or substantially similar to the original "work."[104] The European Court of Justice confirmed in *Infopaq I* that, in such cases, the question is whether those copied, isolated parts of the whole original work sufficiently meet the originality threshold to count as original works themselves.[105] The test, therefore, is not about copying a substantial part, but about copying a sufficiently original part of the work.[106] The substantial similarity test in the U.K., before *Infopaq I*, has previously focused on a qualitative, not quantitative, assessment, which will vary depending on the nature of the copyrighted work and of the infringing work.[107] A compilation or dissection approach to individual features has been rejected in that regard.[108] Subsequent cases in the U.K. have confirmed that the infringement analysis in *Infopaq I* is the correct one.[109] In the United States, the Copyright Act of 1909 initially stated that copyright in a work only extended to "the copyrightable component parts of the work."[110] This wording was removed in the 1976 Act, but its removal was meant to merely clarify existing law rather than to change the established standard of originality.[111] The premise is therefore the same in the United States as it is in the European Union and the U.K.—only the copyrightable aspects of a work are protected against copyright infringement.[112] However,

[104] Sganga, *supra* note 98, at 7.

[105] *Infopaq I*, 2009 E.C.R. ¶¶ 47-51 (holding that "an act occurring during a data capture process, which consists of storing an extract of a protected work comprising 11 words and printing out that extract, is such as to come within the concept of reproduction in part . . . if the elements thus reproduced are the expression of the intellectual creation of their author"). More recently, the Court also confirmed in *Pelham* that "the reproduction by a user of a sound sample, even if very short, of a phonogram must, in principle, be regarded as a reproduction 'in part' of that phonogram within the meaning of the provision, and that such a reproduction therefore falls within the exclusive right granted to the producer of such a phonogram under that provision." *See* Case C-476/17, Pelham GmbH v Ralf Hütter, ECLI:EU:C:2019:624, ¶ 29 (July 29, 2019).

[106] *See* Eleonora Rosati, *What does the European Commission Make of the EU Copyright Acquis when it Pleads before the CJEU? The Legal Service's Observations in Digital/Online Cases*, 45 EUROPEAN L. REV. 67, 71 (2020) (citing the European Commission's Observations in C-406/10, SAS Institute v. World Programming Ltd., ECLI:EU:C:2012:259, ¶¶ 112-13).

[107] *See* Ladbroke (Football) Ltd. v. William Hill (Football) Ltd. [1964] 1 WLR 273, 276; Hawkes & Sons Ltd. v Paramount Film Servs. Ltd. [1934] 1 Ch. 593.

[108] *See* Designers Guild v. Russell Williams (Textiles) Ltd. [2000] UKHL 58 (Lord Millett) (holding that substantiality "is a matter of impression, for whether the part taken is substantial must be determined by its quality rather than its quantity. It depends upon its importance to the copyright work. It does not depend upon its importance to the defendants' work, as I have already pointed out."). In contrast, a mere visual comparison between the works, or compiling or dissecting the individual features of the works is not relevant to substantiality. *Id.* (Lord Hoffmann).

[109] SAS Inst. Inc v World Programming Ltd [2010] EWHC (Ch) 1829 [244].

[110] Copyright Act of 1909, 17 U.S.C. § 3 (repealed 1976).

[111] Feist Publ'ns, Inc. v. Rural Tel. Serv. Co., 499 U.S. 340, 355 (1991).

[112] *Id.* at 361 (phrasing the relevant infringement question, when only a part of a work was copied, as whether that part was "original").

different circuit courts have diverged on the details of assessing substantial similarity. The Second Circuit analyzes substantial similarity by first filtering out similarities that result from unprotectable aspects of the original work, then examining the "total concept and feel, theme, characters, plot, sequence, pace and setting" of the similarities that remain to determine whether the similarities rise to the level of "substantial."[113] The Ninth Circuit, on the other hand, breaks down the analysis into two separate "extrinsic" and "intrinsic" tests.[114] The "extrinsic test" is an objective comparison of specific, protectable expressive elements.[115] The "intrinsic test" is a "subjective comparison that focuses on 'whether the ordinary, reasonable audience'"[116]

Therefore, the reproduction of original works for the purpose of creating a dataset, with means such as downloading, storing, or transferring copies from online sources, will infringe the copyright in those works.[117] There is no requirement in copyright law that a work, if reproduced, must be capable of being readily understood or accessed in order to infringe. This means that even data that is transformed into a different format, including an encrypted format, will still infringe copyright. Once the data has been initially copied to form part of an original dataset, the next steps in the data lifecycle, as discussed,[118] involve pre-processing and transforming the data. Some form of manipulation of the data will often also have occurred at the time of extracting the data from the original source, particularly if it was scraped from the internet.[119] This means that the training dataset will rarely be an identical representation of the original data, but often it will be substantially similar in content and therefore liable for copyright infringement. However, generating synthetic data may involve making more significant changes to the data.[120] The same infringement test applies to synthetic data, and if it is not substantially similar to the original work, it will not infringe on any copyright.

Even assuming that specific copyrighted works have been copied in a dataset, a common difficulty with establishing infringement lies in evidence. Datasets are massive in size, and data may have been processed, structured, or formatted in a way that it is no longer an exact replica of the original data, making it more difficult

[113] Williams v. Crichton, 84 F.3d 581, 588 (2d Cir. 1996).

[114] *E.g.*, Cavalier v. Random House, Inc., 297 F.3d 815, 822 (9th Cir. 2002); Antonick v. Electronic Arts, Inc., 841 F.3d 1062, 1065-66 (9th Cir. 2016).

[115] Benay v. Warner Bros. Ent. Inc., 607 F.3d 620, 624 (9th Cir. 2010).

[116] *Id.*

[117] *See* Directive 2019/790, recital 8, 2019 O.J. (L 130) 92, 93-94 (EU) (hereinafter Copyright Directive) (stating that "[i]n certain instances, text and data mining can involve acts protected by copyright, by the sui generis database right or by both, in particular, the reproduction of works or other subject matter, the extraction of contents from a database or both which occur for example when the data are normalised in the process of text and data mining. Where no exception or limitation applies, an authorisation to undertake such acts is required from rightholders.").

[118] *See supra* Section II.

[119] *Id.*

[120] Manasi Thonte et al., *Technical Review on Synthetic Data Generation*, 9 GRADIVA REV. J. 100 (2023).

to detect.[121] While it can still infringe others' copyright for being substantially similar, if it is not identical, it is like finding a needle in a haystack unless specialized algorithms are employed for detecting the data.[122] Obtaining access to the contents of training datasets, if they are not public, is another critical issue, and there is an overall lack of transparency surrounding data sources among AI developers as highlighted above. The issue of proving infringement is, however, commonplace for copyright cases in general, where there is often a dispute as to whether actual copying has occurred. Many jurisdictions resolve that by establishing a presumption of copying if the plaintiff can prove that the defendant had access to the work at issue,[123] or if there is substantial similarity between the original elements of the infringed and alleged infringing works.[124] The latter option is not particularly helpful in the context of AI, if the infringing work cannot be readily found within a large dataset. However, if it can be proven that the work was publicly available on the internet, which is likely to have become part of a dataset,[125] then that may suffice. To the extent the model can also reproduce, in whole or in part the "memorized" content of the original work, it would further support a claim that there has been access.[126]

B. Exceptions and Limitations That Apply to Data Reproduction for Model Training

In the vast majority of cases, because digital copies of copyrighted works will have been made at some point in the data supply chain, there will be acts of reproduction, even if the original data has subsequently been varied or otherwise manipulated. Those acts of reproduction will infringe copyright unless exempted

[121] Lee et al., *supra* note 7, at 86.

[122] *Id.*

[123] Skidmore v. Zeppelin, 952 F.3d 1051, 1067-69 (9th Cir. 2020) (summarizing case law in the United States for proving access for a copyright infringement claim); Sheeran v Chokri [2022] EWHC (Ch) 827 [26] (United Kingdom) (holding that "[i]rrespective of where the burden lies, infringement requires there to have been actual copying, which necessarily entails that the alleged infringer not only had access to the original work, but actually saw or heard it"). This was also proposed by the Agency for Cultural Affairs of Japan in a recent policy paper, in the context of large datasets used for training AI models. *See* Japan Copyright Off., *General Understanding on AI and Copyright in Japan*, AGENCY FOR CULTURAL AFF. OF JAPAN 13 (May 2024), https://www.bunka.go.jp/english/policy/copyright/pdf/94055801_01.pdf [https://perma.cc/BS3Z-8RW9] (stating that where it is uncertain whether a particular copyrighted material is used in the AI training data, "dependency" (and, therefore, also copying) will be presumed if the copyright holder can prove that the AI user had access to the existing copyrighted work, *or* if the AI-generated material has a high degree of similarity with the work).

[124] Alternatively, if there is a striking similarity between the infringed and alleged infringing works, then access may be inferred on the facts. *See* Three Boys Music Corp. v. Bolton, 212 F.3d 477, 486 (9th Cir. 2000) (citing Granite Music Corp. v. United Artists Corp., 532 F.2d 718, 721 (9th Cir. 1976)). The defendant may then rebut the presumption through proof of independent creation. *Id.*

[125] As previously discussed, a recent comprehensive review showed that there are hundreds of datasets currently available for use by LLMs. *See* Liu et al., *supra* note 20.

[126] The question of whether generative AI models "memorize" their training data is discussed separately below. *See infra* Section III.C.1.

by statutory exceptions or limitations.[127] There seems to be no dispute that without access to copyrighted materials, we would not have witnessed the significant development in generative AI models seen in recent years. This access has enabled significant strides in machine learning performance, allowing models to learn from vast and diverse datasets to improve their outputs in tasks like text generation and image and video synthesis. What is disputed, however, is whether AI developers would have needed authorization from rightholders to train their models on protected content, because their infringing acts fall under statutory exceptions or limitations. Whether acts of data reproduction for model training can be permitted under copyright law is currently pending resolution in lawsuits across the world and has received significant scholarly attention which this Article will not attempt to repeat or replace.[128] Instead, the discussion below will focus on how major copyright regimes and markets such as the United States, EU and the U.K., and others like Japan, Australia, Singapore, and India, differ critically in their treatment of text and data mining and what counts as infringing. Clearly, training generative AI models is not going to be without legal risks anywhere in the world. Yet, as it will be explored in more detail below, it is to be expected that those risks will be more or less pressing in different countries, which both developers and rightholders need to be wary of moving forward.

The right to reproduction is not subject to any harmonized exceptions and limitations in international copyright law. Instead, Article 9(2) of the Berne Convention provides that it shall be a matter of national law to permit the reproduction of works "in certain special cases," provided that such reproduction does not conflict with a "normal exploitation of the work" and does not "unreasonably prejudice the legitimate interests of the author."[129] The same rule can also be found in Article 13 of the TRIPS Agreement,[130] and is commonly known as the three-step test. The three-step test sets a threshold for how far-reaching statutory exceptions or limitations can be. If a signatory state to the Berne Convention violates the three step-test, it can be brought before the World Trade Organization's ("WTO") dispute settlement system. The recommendations and

[127] *See also* EU AI Act, *supra* note 11, at Recital 105 ("Text and data mining techniques may be used extensively in this context for the retrieval and analysis of such content, which may be protected by copyright and related rights. Any use of copyright protected content requires the authorisation of the rightsholder concerned unless relevant copyright exceptions and limitations apply.").

[128] *See* references cited in *supra* notes 4 and 7.

[129] The same wording can also be found in Article 5(5) of the Infosoc Directive, *supra* note 97. The Copyright Directive, *supra* note 117, which is discussed in *infra* Section III.B.2, does not contain the three-step test in the Directive itself, however it refers to the same test in Recital 6.

[130] The Agreement on Trade-Related Aspects of Intellectual Property Rights art. 3, Apr. 15, 1994, 1869 U.N.T.S. 299 [hereinafter TRIPS Agreement].

rulings of the WTO panel become legally binding after being adopted, through reverse consensus,[131] by the dispute settlement body ("DSB").[132]

To date, the WTO has only interpreted the three-step test once, in a dispute in 2000 between the United States and the European Communities.[133] In that case, the WTO interpreted the first step of the test by drawing a significant distinction between "certain" and "special" cases to mean that a statutory "exception and limitation should be narrow in a quantitative as well as a qualitative sense."[134] The second step, "normal exploitation of the work," was interpreted as referring to exploitation that "currently generates significant or tangible revenue," as well as exploitation that, "with a certain degree of likelihood and plausibility, could acquire considerable economic or practical importance."[135] The third and final step, "unreasonably prejudices the legitimate interests of the author," was interpreted to refer only to "legitimate" interests, which will be context-dependent on what is lawful.[136] Furthermore, only "unreasonable" prejudice was deemed unacceptable, which was thought to be the case where an exception or limitation causes or has the potential to cause an unreasonable loss of income to the copyright owner.[137] The WTO panel's interpretation of the three-step test has been widely criticized as unduly narrow and restrictive. A large group of leading scholars issued a joint declaration in 2008 in favor of a balanced interpretation of the three-step test,[138] which argued that the WTO interpretation put too much emphasis on the "public interest" at the cost of rightholders.[139] Another criticism was that the three-step test requires a comprehensive overall assessment, rather than a formulaic step-by-step application, and that the introduction of exceptions or limitations should preclude the payment of compensation below the market rate.[140] An alternative interpretation to the three-step test has also been put forth by Professor Bernt Hugenholtz and Professor Ruth Okediji, who argued that exceptions and limitations that "(1) are not overly broad, (2) do not rob right holders of a real or potential

[131] This means that the DSB which is formed must approve the decision *unless* there is consensus against it. *See* Understanding on Rules and Procedures Governing the Settlement of Disputes, art 16.4, Apr. 15, 1994, 1869 U.N.T.S. 401 (the main WTO agreement on settling disputes between signatory states).

[132] WORLD TRADE ORGANIZATION, A HANDBOOK ON THE WTO DISPUTE SETTLEMENT SYSTEM 61, 88 (2004); Peter Van den Bossche, *The TRIPS Agreement and WTO Dispute Settlement: Past, Present and Future* 5-6 (World Trade Inst., Working Paper No. 02/2020).

[133] Panel report, *United States—Section 110(5) of the US Copyright Act*, WTO Doc. WT/DS160/R (adopted July 27, 2000).

[134] *Id.* ¶ 6.109.

[135] *Id.* ¶ 6.180.

[136] *See id.* ¶ 6.224.

[137] *Id.* ¶ 6.229.

[138] Christopher Geiger et al., *Declaration on a Balanced Interpretation of the "Three-Step Test" in Copyright Law*, 1 JIPITEC, 119-20 (2010).

[139] *Id.* at 119.

[140] *Id.* at 119-120.

source of income that is substantive, and (3) do not do disproportional harm to the right holders, will pass the test."[141]

1. United States

In the United States, the copyright infringement question revolves around whether data reproduction for training amounts to fair use, which is subject to multiple ongoing litigations.[142] Fair use is a multi-factor analysis, which involves assessing:

> (1) the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes;
> (2) the nature of the copyrighted work;
> (3) the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and
> (4) the effect of the use upon the potential market for or value of the copyrighted work.[143]

The first fair use factor, also known as the transformative factor, is met where the infringing work "alter[s] the first [work] with new expression, meaning, or message."[144] The transformative factor was recently discussed by the U.S. Supreme Court in *Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith*.[145] Justice Sonia Sotomayor, writing for the majority, contended that aesthetic or expressive changes do not necessarily make a work transformative. Rather, the central question is whether the new use serves a purpose "sufficiently distinct" from the original.[146] In this regard, the Second Circuit has previously found that secondary use of a work will be transformative where it results in the "creation of new information, new aesthetics, new insights and understandings."[147] There is a reasonable argument that reproductions involved in training datasets are transformative in nature if the data is subsequently used to create new materials that differ from their original sources.[148] This would be the case where expressive materials are used to generate non-expressive content, or where there is little resemblance between generated content and original sources. Other models may, however, be minimally transformative if they are used to create content that highly resembles or even verbatim copies significant portions of the original sources in outputs.[149] The end

[141] P. Bernt Hugenholtz & Ruth L. Okediji, *Conceiving an International Instrument on Limitations and Exceptions to Copyright* 3 (Amsterdam L. Sch., Research Paper No. 2012-43, 2008), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2017629 [https://perma.cc/T6VQ-UGTM].

[142] *See* references to cases cited in *supra* note 4.

[143] 17 U.S.C. § 107.

[144] Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 579 (1994).

[145] Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 598 U.S. 508 (2023).

[146] *Id.* at 550.

[147] Castle Rock Ent. v. Carol Publ'g Group, 150 F.3d 132, 142 (2d Cir. 1998) (citing Pierre N. Leval, *Toward a Fair Use Standard*, 103 HARV. L. REV. 1105, 1111 (1990)).

[148] Quang, *supra* note 7, at 1416-17; Lee et al., *supra* note 7, at 105-06.

[149] Lee et al., *supra* note 7, at 106.

use of the model is relevant for assessing whether there is a transformative purpose and its degree. Many generative AI models, which are general in nature and allow users to create any type of content based on a vast data corpus, will likely count as transformative or even highly transformative.

The second fair use factor, the nature of the copyrighted work, may point in either direction depending on what training data is used. Some training data may be highly expressive and original, whereas other data may be informational or unpublished. The third and fourth factors largely go against AI companies' assertion of fair use. Generative AI models, in particular LLMs, necessarily rely on large amounts of data and have been shown to be capable of rendering expressive works that are similar to original sources used. Some models permit almost verbatim reproduction of segments of training data through specific prompts,[150] although developers recognize that this is a problem and are attempting to prevent it from happening.[151] In common use cases, generative AI models such as ChatGPT or DALL-E will typically not reproduce verbatim copies of original sources. However, there is little doubt that they can still reduce demand for new, original copyrighted materials produced by authors. The generated content may be either "good enough" without the need for any human author, or the generated content may be used to make content creation more efficient, reducing the demand for authors.[152] Generated AI works can serve the same purpose and use case as the underlying, original works. For example:

- A person cannot afford a print, let alone an original, by their favorite landscape artist, Baykalova. They use an AI to generate "a new Baykalova-style landscape art piece." The AI produces an art piece with a similar style, which the person uses instead of purchasing an original print.
- An art dealer decides to sell AI-generated digital art in the style of Picasso. Although not a replica, the art competes with Picasso's original works which are not yet in the public domain and could reduce the demand for his authentic pieces.
- A media company typically requests services from external illustrators on a contract basis. The illustrators make various illustrations of landscapes, objects, etc. However, the company now decides to use an AI to produce illustrations, which are similar to the work which otherwise would be produced by human illustrators. Instead of instructing a group of illustrators, they now write detailed

[150] *See* Exhibit J, N.Y. Times v. Microsoft Corp., No. 23-cv-11195 (S.D.N.Y. filed Dec. 27, 2023) (showcasing how ChatGPT-4 can output near verbatim excerpts of copyrighted source materials).

[151] *OpenAI and journalism*, OPENAI (Jan. 8, 2024), https://openai.com/index/openai-and-journalism [https://perma.cc/R424-NBU9] (describing memorization as a "a rare failure of the learning process that [OpenAI] are continually making progress on").

[152] That AI-generated works may reduce the demand of original copyrighted works that are used as training data has been raised in several of the ongoing copyright lawsuits. *See* cases cited *infra* note 440.

prompts that describe what they want produced. Although it is not a replica of the work that would have been produced by human illustrators, it satisfies the company's needs.

- A news organization decides to use AI to generate articles on routine topics like sports scores, weather updates, and stock market reports. This reduces the need for human journalists to write these articles, potentially leading to layoffs or reduced hiring.

That generated AI works can reduce demand for the works produced by original authors is further inferred from the fact that authors across the world have gone on strikes. For example, in May 2023, screenwriters in the United States went on a nationwide strike, demanding that generative AI models only be used as tools that facilitate script writing rather than replacements of writers.[153] There is also a commercial market for training data that is rapidly growing. For example, one company, Hazy, recently launched its public marketplace for synthetic data, allowing data producers to publish generator models.[154] OpenAI and other AI developers have also entered into licensing agreements for training purposes with a growing number of companies. Recently, OpenAI has struck deals with content producers and aggregators like Associated Press, Axel Springer, Financial Times, Reddit, Vox Media and Shutterstock, among others.[155] On the one hand, the fact that training data licensing is happening could further support that there should be no fair use argument to reproduce data for model training. If licensing deals are made, then there is arguably a licensing market that rightholders lose out on if their content is used without permission as data. On the other hand, this may just be a pragmatic way for AI developers to reduce their risks amidst current copyright infringement uncertainties, to get access to data that is not publicly available from online scraping, or to get access to such data in a more effective way.

The heart of the copyright infringement issue for AI-generated works is that copyright protects specific and original expressions of ideas, genres, and styles, but not ideas, genres, or styles themselves. This idea-expression dichotomy is based on the premise that prohibiting the free use of ideas will run against the policy

[153] Alissa Wilkinson, *The Looming Threat of AI to Hollywood, and Why it Should Matter to You*, VOX (May 2, 2023), https://www.vox.com/culture/23700519/writers-strike-ai-2023-wga [https://perma.cc/U2X4-JRXW]. That eventually resulted in a union labor agreement, requiring that AI is not used to write source literary material and that writers are free to choose whether to use AI or not to facilitate their writing. That same agreement stated that the Writers Guild of America West reserved the right to "assert that exploitation of writers' material to train AI is prohibited by MBA or other law." *See Summary of the 2023 WGA MBA*, WRITERS GUILD OF AMERICA WEST (Oct. 24, 2023) https://www.wga.org/contracts/contracts/mba/summary-of-the-2023-wga-mba [https://perma.cc/QC2G-2ZEC].

[154] Harry Keen, *Launching our Public Synthetic Data Marketplace*, HAZY (June 5, 2024), https://hazy.com/resources/2024/06/05/public-synthetic-data-marketplace#whyhazy [https://perma.cc/KTE4-HFRR].

[155] *AI content licensing deals: Where OpenAI, Microsoft, Google, and others see opportunity*, CBINSIGHTS (July 19, 2024), https://www.cbinsights.com/research/ai-content-licensing-deals/ [https://perma.cc/RS3R-N8SD].

objective of copyright law to encourage creativity.[156] Ideas, genres, styles, or so-called *scènes à faire*, should not be monopolized themselves and may be used as "common building blocks" to inspire further creative expressions that build upon them.[157] U.S. courts have long upheld the *scènes à faire* doctrine, refusing to extend copyright protection to similarities between works at a higher or more generalized level.[158] If the same logic is applied to AI-generated works, this could mean that AI-generated works themselves will in many cases not infringe on the copyright of the original work, simply because they will not meet the substantial similarity test. However, and importantly, just because the end use of the generative AI model is not infringing does *not* necessarily suggest that the fourth fair use factor is met with respect to infringing reproductions of training data. A fair use cannot "usurp the market of the original work."[159] Specifically, the U.S. Supreme Court has said, "to negate fair use one need only show that if the challenged use should become widespread, it would adversely affect the potential market for the copyrighted work."[160]

Whether AI-generated works, whether or not they infringe themselves, meet this test or not will ultimately have to be tried by the courts. "Memorized" versions of training data that are generated as near-replicas will infringe when prompted by the end user, as any other work that is identical or substantially similar. However, as will be discussed separately below, it seems to be the exception rather than the norm that training data gets "memorized" in datasets, and it will only be in more rare circumstances that AI-generated works will count as infringing derivative

[156] MELVILLE B. NIMMER & DAVID NIMMER, NIMMER ON COPYRIGHT § 2.03[D] (2014).

[157] Hartford House, Ltd. v. Hallmark Cards, Inc., 846 F.2d 1268, 1274 (10th Cir. 1988) (holding that copyright does not confer "exclusive rights in an artistic style or in some concept, idea, or theme of expression. Rather it is the . . . specific artistic expression . . . that is being protected."); Olson v. National Broadcasting Co., 855 F.2d 1446, 1451 (9th Cir. 1988) (holding that the script for a television show, which had similarities as which were "common to the genre," and therefore could not infringe as substantially similar to another show). *See also* Omri Rachum-Twaig, *A Genre Theory of Copyright*, 33 SANTA CLARA HIGH TECH. L.J. 34, 64-81 (2016).

[158] Gates Rubber Co. v. Bando Indus., Ltd., 9 F.3d 823, 838 (10th Cir. 1993) (holding that "[u]nder the scenes a faire doctrine, we deny protection to those expressions that are standard, stock, or common to a particular topic or that necessarily follow from a common theme or setting. . . . Granting copyright protection to the necessary incidents of an idea would effectively afford a monopoly to the first programmer to express those ideas."); Olson v. National Broadcasting Co., 855 F.2d 1446, 1451 (9th Cir., 1988) (holding that there was no infringement in a television series script, where the two shows "emphasize action and lack identifiable themes. Both shows may be broadly described as comic, and they therefore have similar moods. Both works are quickly paced. However, these similarities are common to the genre of action-adventure television series and movies and therefore do not demonstrate substantial similarity").

[159] NXIVM Corp. v. Ross Inst., 364 F.3d 471, 482 (2d Cir. 2004).

[160] Harper & Row Publishers, Inc. v. Nation Enters., 471 U.S. 539, 568 (1985) (quotation marks omitted). In *Goldsmith*, Justice Neil Gorsuch, concurring with the majority, described the fourth fair use factor as asking, "whether consumers treat a challenged use 'as a market replacement' for a copyrighted work or a market complement that does not impair demand for the original." Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 598 U.S. 508, 555 (2023) (Gorsuch, J., concurring).

works.[161] If that is the case, then courts will need to tackle the tricky question whether the *scènes à faire* doctrine should have an impact on what market effects are deemed relevant for assessing whether the fourth fair use factor is met.

In sum, it is likely that some generative AI models will count as fair use, depending on what data is used and how and what guardrails exist to prevent verbatim copying. However, models that can be used without difficulty to generate near-substantially, or substantially similar representations of copyrighted materials, which directly compete with authors, are more likely to be infringing and not fair use. Generative AI models, which can generate identical representations, are obviously at an even greater infringement risk, but the big question will be *who* will be liable for those direct copyright infringements. It is clear that end-users will be primarily liable, but it is not clear whether dataset creators and model providers will be secondarily liable.[162]

Furthermore, not only developers and providers of generative AI models, but also independent dataset creators, including web scrapers, will directly infringe copyright when reproducing any copyrighted works in the dataset itself.[163] It is an open question at this stage whether dataset creators would be better positioned, compared to AI developers and providers, in a fair use case. For example, some independent dataset creators or web scrapers are non-profit organizations,[164] which might be looked more favorably upon when assessing the first fair use factor.[165] The fourth fair use factor might also be considered differently, if the dataset can further be used for substantial non-infringing purposes, which would be the case if "memorization" of training data is only a rare technical occurrence.

[161] *See infra* Sections III.C.1, VIII.B (discussing references that suggest that extractable memorization rates can be as low as 1-2% in several LLMs or diffusion-based models, whereas in more specialized training datasets the memorization rate can be much higher). *See also infra* Section VIII.A (discussing whether generated output could count as derivative works).

[162] For a discussion regarding whether these questions, see *infra* Section VIII.B (suggesting that it will be the exception, rather than the norm, that dataset creators and model providers will become secondarily liable for copyright infringement, if the model is capable of substantial non-infringing use and there is no actual knowledge of specific infringing acts, in the form of "memorized" training data).

[163] Lee et al., *supra* note 7, at 99.

[164] *See* Andy Baio, *AI Data Laundering: How Academic and Nonprofit Researchers Shield Tech Companies from Accountability*, WAXY (Sept. 30, 2022), https://waxy.org/2022/09/ai-data-laundering-how-academic-and-nonprofit-researchers-shield-tech-companies-from-accountability/ [https://perma.cc/K9VX-YSQ9].

[165] On the other hand, the U.S. Supreme Court has found that "[t]he crux of the profit/nonprofit distinction is not whether the sole motive of the use is monetary gain, but whether the user stands to profit from exploitation of the copyrighted material without paying the customary price." *See* Harper & Row Publishers, Inc. v. Nation Enters., 471 U.S. 539, 562 (1985). This could be understood as meaning that, if a license typically would have been required for using the work, then the fact the use occurred in a non-profit setting should not make it immune from infringement.

2. European Union

i. Background of the Text and Data Mining Exception

The copyright framework in the EU operates differently from the United States, with a long list of specific exceptions and limitations setting out in detail what use is permissible. In the recently enacted Directive 2019/790 (the "Copyright Directive"), Article 4(1) requires Member States to provide for an exception or limitation for reproductions and extractions of lawfully accessible works and other subject matter for the purposes of text and data mining. Text and data mining is broadly defined in Article 2(2) as any automated analytical technique aimed at analyzing text and data in digital form in order to generate information, which includes but is not limited to patterns, trends and correlations. This would include extracting data in datasets for automated processing.[166] The basic premise, therefore, is that publicly available online material may be reproduced for text and data mining purposes. Recital 10 of the Directive justified this on the basis that "[a]s research is increasingly carried out with the assistance of digital technology, there is a risk that the Union's competitive position as a research area will suffer, unless steps are taken to address the legal uncertainty concerning text and data mining."

Text and data mining is a transactional problem at a massive scale.[167] To undertake text and data mining activities, a large body of copyrighted material must be obtained. From a legal point of view, these collections must be reviewed on a case-by-case basis to sift through works that are not copyrightable and works that are copyrightable. For works that are copyrightable, the person undertaking text and data mining must then identify who is the relevant rightholder, track them down, and obtain their permission to reproduce their works in the dataset, whether for free or for a license fee.[168] The transaction costs of undertaking this very extensive review process in a large dataset may far exceed the cost of obtaining

[166] *See Commission Staff Working Document Impact Assessment on the Modernisation of EU Copyright Rules*, at 104, COM (2016) 593 final (Sept. 14, 2016) [hereinafter *Commission Impact Assessment*] (describing text and data mining as "a term commonly used to describe the automated processing ('machine reading') of large volumes of text and data to uncover new knowledge or insights"). For a legal study which was commissioned by the European Commission in preparation for the Copyright Directive, *supra* note 117, see Jean-Paul Triaille et al., *Study on the legal framework of text and data mining (TDM)*, at 28 (Mar. 2014), https://op.europa.eu/s/zYaH (describing text and data mining as including the following steps of "[i]ndividual content is extracted from outside sources," "[c]ontent is, when necessary, transformed to fit operational needs," "[c]ontent is loaded into a data set, repository or collection," "[d]ata miners gain access to the data and the mining (analysis) tools are applied to the data set," and "[n]ew knowledge is created as a result of the analysis").

[167] Gregor Langus, Damien Neven & Gareth Shier, *Assessing the Economic Impacts of Adapting Certain Limitations and Exceptions to Copyright and Related Rights in the EU*, at 43, 75 (October 2013), https://op.europa.eu/s/zYaG.

[168] *See supra* Section III.A.

each individual license.[169] In the early and mid-2010s, when the Copyright Directive was first considered by the European Commission, this was particularly problematic for research institutions who lacked the means to undertake extensive and expensive copyright compliance checks for their text and data mining activities for the purpose of conducting research.[170] Reducing these transaction costs for research institutions was therefore cited as one of the fundamental policy reasons for a text and data mining exception.[171]

Another important reason behind the text and data mining exception was to address the legal uncertainty concerning text and data mining.[172] In the underlying impact assessment for the Copyright Directive, the European Commission explained that researchers face difficulties regarding protected content to which they already have lawful access to on the basis of subscriptions purchased by their libraries or institutions.[173] In particular, it was stated that "[s]ubscriptions to scientific publications may currently include or not the authorization to perform [text and data mining], prohibit it altogether, or leave it unclear."[174] Recitals 9 and 18 of the Copyright Directive, as enacted, also explain that not all reproductions made during text and data mining will fall under existing exceptions and limitations for temporary acts of reproduction. Indeed, as discussed above,[175] it will be the exception, rather than the norm, that copies of data forming part of a dataset are only transient or temporary copies.[176] It was against this background that the current wording of the text and data mining exception was first proposed and eventually formulated in the Copyright Directive. Yet at this time, AI development was far from at the level of maturity and scale as it is today. AI is not mentioned a single

[169] The issue of transaction costs was highlighted in the impact assessment conducted by the European Commission before considering its proposal for the draft Copyright Directive, *supra* note 117. *See Commission Impact Assessment*, *supra* note 166, at 105-06. An economic study, referred to in the same impact assessment, also describes these transaction costs in greater detail. *See* Diane McDonald & Ursula Kelly, *Value and benefits of text mining,* JISK (Mar. 14, 2012), https://www.jisc.ac.uk/reports/value-and-benefits-of-text-mining [https://perma.cc/AS67-Z5HN].

[170] *See Commission Impact Assessment*, *supra* note 166, at 104-06 (describing how text and data mining is a powerful "scientific research tool" and highlighting how "researchers" in particular consider copyright to be problematic for their operations).

[171] *Id.* at 105, 111-13, 116, 118-19 (repeatedly mentioning the "reduction of transaction costs" for researchers carrying out text and data mining activities).

[172] Copyright Directive, *supra* note 117, at Recital 11 (stating that "[t]he legal uncertainty concerning text and data mining should be addressed by providing for a mandatory exception for universities and other research organisations, as well as for cultural heritage institutions, to the exclusive right of reproduction and to the right to prevent extraction from a database.").

[173] *See Commission Impact Assessment*, *supra* note 166, at 105.

[174] *Id.*

[175] *See supra* Section III.A.

[176] *See also* Triaille et al., *supra* note 166, at 46 (stating that "[i]t is further unlikely that a temporary copy used to mine data is transient, the work mostly being available for a certain period of time to be transformed, loaded and/or analyzed. The extraction will not be transient if the removal of the transient copy does not happen automatically and depends on a human intervention.").

time in the Copyright Directive itself nor in the impact assessment conducted by the European Commission.[177]

In the European Commission's initial proposal for the Copyright Directive, a text and data mining exception was only provided for research organizations for the purpose of carrying out scientific research,[178] which later became Article 3(1) of the adopted Directive. The proposal was discussed extensively between Member States. The research-focused text and data mining exception was eventually joined by additional exception and limitation, what is now Article 4(1) of the Copyright Directive, which is not limited to any particular type of organization or purpose. The new text and data mining exception in Article 4(1) was initially set out as optional for Member States to introduce, which was supported by the majority of the delegations.[179] During the negotiations, it was discussed whether this new exception should be more limited in its scope—for example, by covering only temporary copies of works and other subject matters that have been made freely available to the public online.[180] This was not accepted by the majority of the delegations.[181]

Article 4(1) of the Copyright Directive, as adopted, is not limited to text and data mining activities conducted by research organizations or for the purpose of conducting research. It is also mandatory for Member States to introduce.[182] The compromise that was initially struck when introducing the text and data mining exception in Article 4(1) for commercial entities and non-research purposes was to make it conditional that rightholders have not expressly reserved their rights in this regard. Specifically, Article 4(3) provides that the exception in Article 4(1) shall only apply on the condition that rightholders have not "expressly reserved" their rights in an "appropriate manner," such as machine-readable means in the case of content made publicly available online. Therefore, rightholders can "opt-out" of the data mining exception within the EU, if they so wish. This narrows down the scope

[177] However, the EU AI Act specifically commented that the text and data mining exception in the Copyright Directive applies to the AI context. EU AI Act, *supra* note 11, at Recital 105 ("Directive (EU) 2019/790 introduced exceptions and limitations allowing reproductions and extractions of works or other subject matter, for the purpose of text and data mining, under certain conditions. . . . Where the rights to opt out has been expressly reserved in an appropriate manner, providers of general-purpose AI models need to obtain an authorisation from rightsholders if they want to carry out text and data mining over such works.").

[178] *Proposal for a Directive on Copyright in the Digital Single Market*, at art. 3.1, COM (2016) 593 final (Sept. 14, 2016).

[179] *See* Estonian Presidency, *Proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market - Presidency compromise proposal (consolidated version) and state of play*, at 2, 4, Doc. 15651/17 (Dec. 13, 2017), https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CONSIL:ST_15651_2017_INIT [https://perma.cc/L9LX-XT7Y].

[180] *See* Estonian Presidency, *Proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market - Mandate for Negotiations with the European Parliament*, at 3, Doc. 8145/18 (Apr. 23, 2018), https://data.consilium.europa.eu/doc/document/ST-8145-2018-INIT/en/pdf [https://perma.cc/X7RB-6L2P].

[181] *Id.*

[182] *See* Copyright Directive, *supra* note 117, at art. 4(1) ("Member states *shall* provide for an exception or limitation") (emphasis added).

of the text and data mining exception in Article 4(1) and balances it with the interests of rightholders, who remain in control of the extent their works can be used in datasets. Indeed, this narrower scope could be deemed necessary for the exception and limitation to meet the three-step test in the Berne Convention and TRIPS Agreement, as discussed.[183] The fact that rightholders may reserve their rights means that the text and data mining is unlikely to conflict with the normal exploitation of their works or unreasonably prejudices their legitimate interests.[184]

ii. When Rightholders Will Have "Opted Out"

It is not specified how this opting-out mechanism works in practice, and there are several questions that remain open. It is not fully clear, for example, whether rightholders' opt-out should have a retroactive effect for works that have been previously published on the internet, which may have already been subject to text and data mining. It is also not fully clear what type of "reservation" is sufficient, but it appears to be sufficient, for example, to include a statement that data mining is not a permitted activity in the website's general terms and conditions[185] or in robots.txt.[186] In a recent case in Germany against LAION, which scrapes websites to build datasets, the District Court of Hamburg found at a hearing that a passage in a subsection of its website's terms of conditions constituted a valid opt-out within the meaning of Article 4(3) of the Copyright Directive.[187] The website's terms of conditions did not expressly refer to that Article or mention text and data mining. Instead, it broadly stated that "YOU MAY NOT . . . [u]se automated programs, applets, bots or the like to access the Bigstock.com website or any content thereon for any purpose, including, by way of example only, downloading Content, indexing, scraping or caching any content on the website."[188] The Court expressed

[183] *See supra* Section III.B. *See also* Martin Senftleben, *Generative AI and Author Remuneration*, 54 INT'L REV. INTELL. PROP. & COMPETITION L. 1535, 1544-45 (2023).

[184] *Id.* at 1545.

[185] *See* Copyright Directive, *supra* note 117, at Recital 18 ("In the case of content that has been made publicly available online, it should only be considered appropriate to reserve those rights by the use of machine-readable means, *including metadata and terms and conditions of a website or a service*." (emphasis added)). If a text and data mining reservation is narrow in scope, for example for only reserving the rights in respect of some works, then it could become highly impractical for developers to track down the particular works which are affected, and which ones are not, by the reservation.

[186] *Id.* ("*including metadata*" (emphasis added)). Robots.txt is a file which can be used to manage crawling traffic. *See Introduction to robots.txt*, GOOGLE: SEARCH CENTRAL, https://developers.google.com/search/docs/crawling-indexing/robots/intro [https://perma.cc/BM3X-8ZJ4] (last visited October 27, 2024). OpenAI and Anthropic have publicly said that they respect robots.txt and other blocking mechanisms to their specific web crawlers, GPTBot and ClaudeBot. *See supra* note 19.

[187] LG, July 11, 2024, 310 O 227/23, openJur (Ger.), https://openjur.de/u/2495651.ppdf [https://perma.cc/BBA5-RY9T]. *See* Paul Keller, *Machine readable or not? – notes on the hearing in LAION e.v. vs Kneschke*, WOLTERS KLUWER: KLUWER COPYRIGHT BLOG (July 22, 2024), https://copyrightblog.kluweriplaw.com/2024/07/22/machine-readable-or-not-notes-on-the-hearing-in-laion-e-v-vs-kneschke/ [https://perma.cc/X8G2-DJ3J].

[188] Keller, *supra* note 187.

during the hearing that this was a clear opt-out from text and data mining.[189] The parties also debated at the hearing whether the opt-out was deemed "machine-readable" as required by Article 4(3) of the Copyright Directive. LAION argued that the opt-out should be provided in a standardized format, such as a robots.txt file, which can easily be recognized by crawlers and other bots. The plaintiff, on the other hand, argued that digital plain text is sufficiently "machine-readable," and that requiring opt-outs to be made in specific technical formats would be inappropriate, as many authors do not possess such technical knowledge. None of these issues were, unfortunately, discussed in the Court's decision, which was handed down in September 2024.[190]

Even if an opt-out is deemed to have been made effectively on a website; however, it is further not clear who should be responsible for any ambiguities in such website statements. For example, if an author initially states on his or her website, "I reserve my rights under Article 4(3) of the Copyright Directive for all my works,"[191] but does not specify which works those are, then developers are left in a difficult situation to find that out, which would run against the objective of reducing transaction costs. Similar difficulties arise if such statements should be interpreted to refer to all works publicly available on all websites, or if statements like "all rights reserved" are made on a website. A more pragmatic approach would be to narrowly interpret any opt-out statements, which would be consistent with the underlying objective of reducing transaction costs for rights clearance and the fact that the reservation must be made "expressly" and in an "appropriate manner." Otherwise, the benefit of a text and data mining exception for non-research purposes will largely become null and void. If this understanding is correct, this would mean that, if rightholders have expressly reserved their rights with respect to works published on a particular website, and if those same works can be found on another website, which are not subject to the same reservation, then those latter works would, in principle, be free to be mined, so long as they are "lawfully accessible works."

[189] *Id.*

[190] This is because the Court found that LAION's data scraping practice was permissible under the separate research exception in Article 3(1) of the Copyright Directive, *supra* note 117, for which there is no possibility to opt-out. *See* LG, July 11, 2024, 310 O 227/23, openJur (Ger.), https://openjur.de/u/2495651.ppdf [https://perma.cc/BBA5-RY9T].

[191] Recital 18 of the Copyright Directive, *supra* note 117, might be interpreted to permit such broad reservations ("[i]n other cases, it can be appropriate to reserve the rights by other means, such as contractual agreements or a *unilateral declaration*.") (emphasis added). It is not discussed any further, however, what a "unilateral declaration" means. The Swedish government however, when implementing the Directive into national law, has assumed an extremely narrow interpretation of what "lawful access" means. Specifically, it considered that "lawful access" meant access to works by the right holder's consent or as permitted by law, and gave the example that it is possible to procure a license to access works. *See* Proposition [Prop.] 2021/22:278, at 220, Upphovsrätten på den digitala inre marknaden [government bill] (Swed.). If this is the correct interpretation, which seems unlikely, then this would make the text and data mining exception meaningless for most dataset creators and developers who rely on scraping the internet for publicly available content, but without the permission of rightholders.

iii. What Is Meant by "Lawfully Accessible Works"

The Copyright Directive does not expressly define what "lawfully accessible works and other subject matter" mean. However, Recital 18 suggests that they should be equated with content that has been made publicly available online with the permission of the rightholder.[192] The term "lawful access" is also used in the context of the text and data mining exception in Article 3(1), which is commented on in Recital 14. Specifically, Recital 14 explains that "[l]awful access should be understood as covering access to content based on an open access policy or through contractual arrangements between rightholders and research organizations or cultural heritage institutions, such as subscriptions, *or through other lawful means*" (emphasis added). It then goes on to say that "[l]awful access should also cover access to content that is freely available online."[193] This together suggests that "lawfully accessible works" should mean works that are not infringing. Infringing works have, of course, not been made publicly available with the permission of the rightholder. Indeed, the functioning of the text and data mining exception would become out of balance if developers could consciously focus on mining infringing content, to which rightholders have not had the chance to "opt-out" from.

It is not fully clear if circumventing technological protection measures to access works for text and data mining purposes renders such works to have been "unlawfully accessed."[194] Recital 7 seems to suggest that the exceptions and limitations provided in the Copyright Directive apply, notwithstanding any such protection measures.[195] That exceptions and limitations trump any technological

[192] Copyright Directive, *supra* note 117, at Recital 18, ¶ 2 ("This exception or limitation should only apply where the work or other subject matter *is accessed lawfully by the beneficiary, including when it has been made available to the public online*, and insofar as the rightholders have not reserved in an appropriate manner the rights to make reproductions and extractions for text and data mining.") (emphasis added).

[193] In the *Commission Impact Assessment*, *supra* note 166, at 108, it was similarly suggested, but in the context of text and data mining for non-commercial scientific research purposes, that "[l]awful access would cover access to content through authorisation by content owners (e.g. subscriptions to scientific journals) as well as access to publicly available content (e.g. open access content)." *See also* Thomas Margoni & Martin Kretschmer, *A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology*, 71 GRUR INT'L 685, 697 (2022).

[194] Article 6(1) of the Infosoc Directive, *supra* note 97, provides that Member States shall provide adequate legal protection against the circumvention of any effective technological measures, which the person concerned carries out in the knowledge, or with reasonable grounds to know, that he or she is pursuing that objective. Such technological measures are defined in Article 6(3) as "any technology, device or component that, in the normal course of its operation, is designed to prevent or restrict acts, in respect of works or other subject-matter, which are not authorised by the rightholder of any copyright or any right related to copyright as provided for by law or the *sui generis* right provided for in Chapter III of Directive 96/9/EC." *Id.*

[195] Specifically, Recital 7 states that "[t]he protection of technological measures established in Directive 2001/29/EC remains essential to ensure the protection and the effective exercise of the rights granted to authors and to other rightholders under Union law. *Such protection should be maintained while ensuring that the use of technological measures does not prevent the enjoyment of the exceptions and limitations provided for in this Directive.*" (emphasis added). *Id.*

protection measures is consistent with what is the case in the Infosoc Directive. Article 6(4) of the Infosoc Directive provides that, in the absence of voluntary measures taken by rightholders to implement such technological protection measures, Member States shall take steps to ensure that rightholders make available to the beneficiary of an exception or limitation provided for in national law.[196] In practice, this matters little; however, rightholders are still in control of expressly opting-out of any text and data mining activities.

Copyright infringement is a strict liability tort.[197] In principle, if a work that has *not* been made lawfully accessible is included in the training dataset, then developers will infringe on the copyright for that work, regardless of whether rightholders have opted out or not. Practically, this means that AI developers in the EU must carefully vet the training datasets *before* they are reproduced to ensure that they do not contain works that are not "lawfully accessible works." This is easier said than done and raises difficult questions. Content uploaded to a particular website may inadvertently contain infringing materials, which are not "lawfully accessible works." If the broader internet has been crawled or scraped, it is likely that infringing materials have been included in the dataset, unless proven otherwise. On the one hand, it would be convenient to establish a reverse burden of proof in such cases, moving the onus onto the AI developers to prove that the dataset does not contain the individual rightholder's works from an unauthorized source. Rightholders will inevitably lack the same insight into what data has been used by developers, who, at least in theory, have the means to control and vet the data for infringing content before reproducing it for training purposes. On the other hand, the vast quantities of data may make it practically infeasible, or even nearly impossible, for developers to investigate infringing content in closer detail. Copyright infringement is a complex analysis and supposes that there has been copying without permission of the individual rightholder. It is frequently not stated on websites where data come from and whether they have been reproduced with rightholders' permission. It is unclear how developers should tackle this problem. Full copyright compliance in AI development would require substantial investments to ensure compliance by automated means, if it is at all realistic with

[196] This is also confirmed in Recital 7 of the Copyright Directive, *supra* note 117 ("[i]n the absence of voluntary measures, Member States should take appropriate measures in accordance with the first subparagraph of Article 6(4) of Directive 2001/29/EC, including where works and other subject matter are made available to the public through on-demand services."). In one of the legal studies on text and data mining that were submitted to the European Commission before its proposal for the Directive, it was similarly suggested that the circumvention of technological protection measures would *not* render the access unlawful. *See* Triaille et al., *supra* note 166, at 72-74, 112-13.

[197] LIONEL BENTLY & BRAD SHERMAN, INTELLECTUAL PROPERTY LAW 143 (5th ed. 2018); Educational Testing Serv. v. Simon, 95 F. Supp. 2d 1081, 1087 (C.D. Cal. 1999) (holding that copyright infringement "is a strict liability tort"); King Records, Inc. v. Bennett, 438 F. Supp. 2d 812, 852 (M.D. Tenn. 2006) ("[A] general claim for copyright infringement is fundamentally one founded on strict liability."); Faulkner v. Nat'l Geographic Soc'y., 576 F.Supp.2d 609, 613 (S.D.N.Y., 2008) ("Copyright infringement is a strict liability wrong in the sense that a plaintiff need not prove wrongful intent or culpability in order to prevail."). *But, see* Patrick Goold, *Is Copyright Infringement a Strict Liability Tort?*, 30 BERKELEY TECH. L.J. 305 (2015) (arguing that copyright infringement should not be considered a strict liability tort in countries with a fair use doctrine).

current technologies, and it is questionable whether the benefits would ultimately outweigh the costs. Because one of the intentions behind the text and data mining exception was to reduce transaction costs, as discussed, it would be inappropriate to assume that full copyright compliance is necessary when extracting "lawfully accessible works."

A middle-ground between these two extremes would be to introduce a knowledge requirement for AI developers and those using AI tools, meaning they only infringe if they have actual, imputed, or constructive knowledge that the dataset contained infringing content, including those that had been "opted out" from the text and data mining exception.[198] This would be similar to what is the case for hyperlinking. In the landmark *GS Media* case, the European Court of Justice established that someone posting a hyperlink will be liable for copyright infringement. if they know or ought to have known that the linked content was infringing.[199] Where someone is carrying out linking activities for profit, the Court also made it clear that whoever posts such hyperlinks is expected to check to ensure that the work concerned is not "illegally" published on the website to which that hyperlink leads.[200] Consequently, it was held that it must be presumed that posting that hyperlink for profit occurred in full knowledge of the illegal nature of that publication.[201] Introducing a similar presumption of knowledge for commercial AI developers would mean that we go back to square one and require developers to conduct full infringement searches in massive data. A better solution would be to require actual, imputed, or constructive knowledge of infringing material in the dataset but without the presumption. In practice, this could mean that it is to be expected that AI developers exclude sources that are commonly known to contain infringing materials from being part of the dataset. This would be similar to what is the case in Singapore. The Singaporean Copyright Act, discussed further below,[202] permits the mined material to be an infringing copy, if the person conducting the text and data mining activity does not know this, *and* in case the

[198] A similar policy was proposed by the Agency for Cultural Affairs of Japan in its recent policy paper, stating that AI developers or service providers may commit infringement if they have knowledge that the training data comes from websites that contain infringing content. *See* Japan Copyright Off., *supra* note 123, at 11. The position is also similar in Singapore, where Section 244(2)(e) of the Copyright Act 2021, No. 22 of 2021 (Sing.), provides that it is not an infringement to use data for computational data analysis, even if the data contains infringing content, where the person does not know that it is infringing, or could not have reasonably known that it came from a flagrantly infringing online location, or where using of infringing copies is necessary for a prescribed purpose and they are only used to carry out computational data analysis.

[199] Case C-160/15, GS Media BV v. Sanoma Media Netherlands BV, ECLI:EU:C:2016:644, ¶ 49 (Sept. 8, 2016).

[200] *Id.* ¶ 51.

[201] *Id.* For a more detailed analysis of the *GS Media* case and the questions of choice of law and localization of hyperlinking infringements, see Mattias Rättzén, *Cross-Border "Illegal" Linking: Questions of Localization and Choice of Law*, 14 J. INTELL. PROP. L. & PRACT. 539 (2019).

[202] *See infra* Section III.B.4.

copy "is obtained from a flagrantly infringing online location . . . [he or she] does not know and could not reasonably have known that."[203]

iv. When Text and Data Mining Is Conducted for Scientific Research

As mentioned above, Article 3(1) of the Copyright Directive also provides a text and data mining exception but is limited to "research organisations and cultural heritage institutions" and "for the purposes of scientific research." Recital 12 makes it clear that "scientific research" excludes organizations upon which commercial undertakings have a decisive influence by allowing such undertakings to exercise control due to structural situations, such as through the quality of shareholders or members. It is accepted in the same Recital, however, that research organizations do not necessarily have to be related to education or science but can also cover organizations that act on a not-for-profit basis or in the context of a recognized public-interest mission, for example, through public funding. This raises the question of whether non-profit and open data crawl repositories, such as Common Crawl, could be deemed to undertake data mining for the purpose of scientific research. What is unique about these organizations is that they have been set up to hoard massive amounts of data scraped from the internet, which is then made available for use for a variety of different purposes. The significant advantage of being able to rely on Article 3(1) instead of Article 4(1) for text and data mining is that there is no possibility for rightholders to reserve their rights and opt-out. The Hamburg District Court recently found in *Robert Kneschke v LAION e.V.* that LAION, which is a non-profit organization focused on web scraping publicly available sources, was a qualifying research organization within the meaning of Article 3(1) of the Copyright Directive, as implemented into German law. Because LAION's dataset was publicly available for free and intended for research purposes, its web scraping practices were a permitted research activity, for which there is no need to respect rightholders' opt-outs.[204] However, this broad reading of the research text and data mining exception can be criticized on the basis that it disregards the fact that the end use of the training datasets provided can vary from non-commercial to commercial. The Court also disregarded the fact that text and data mining exceptions only exempt acts of reproduction of training data from infringing. Yet a separate act of communication to the public also occurs when training datasets are provided on the internet for others to access.

If it is correct that non-profit organizations could be deemed to carry out "scientific research" when hoarding massive datasets through web scraping or web crawling, then this raises the critical question of whether third party commercial AI developers using such open data repositories could take advantage of that as well. This becomes relevant as it is well-known, as discussed,[205] that AI developers are using publicly available open data repositories to a significant extent when putting

[203] Copyright Act 2021, No. 22 of 2021, § 244(2)(e)(ii) (Sing.).

[204] *See* LG, July 11, 2024, 310 O 227/23, openJur (Ger.), https://openjur.de/u/2495651.ppdf [https://perma.cc/BBA5-RY9T].

[205] *See supra* Section II.

together their training datasets. If such use by commercial AI developers could similarly benefit from Article 3(1), then it would arguably create a loophole in the Copyright Directive that could be exploited to the detriment of rightholders. Instead of respecting opt-outs by rightholders, AI developers could simply turn to an open data repository that was created by another organization. This cannot be what was intended when formulating the rules and might fail to meet the three-step test under the Berne Convention.[206] As discussed above,[207] it is the fact that rightholders may reserve their rights that is likely to ensure compliance with the three-step test, avoiding the fact that text and data mining has a significantly adverse effect on the normal exploitation of their works. There is also a significant licensing market for using training data, which could be circumvented if commercial AI developers rely on Article 3(1) through open data repositories. Furthermore, and importantly, the text and data mining exceptions provided in both Articles 3(1) and 4(1) are triggered when "reproductions and extractions" are made. Commercial AI developers who use open data repositories will need to make copies of the dataset, which will count as infringing reproductions.[208] Because of that, and because "text and data mining" is broadly defined in Article 2(2) to refer to "any automated analytical technique aimed at analyzing text and data in digital form in order to generate information," it is likely that commercial AI developers who use open data repositories will still be deemed to primarily engage in text and data mining themselves. Therefore, it seems highly unlikely that commercial AI developers could become immune from having to respect rightholders' opt-outs by simply using a third-party open data repository, even if that has been created for the purpose of scientific research.[209]

v. The Obligation to Draw up a Summary of Training Data

Regulation 2024/1689 (the "EU AI Act") on artificial intelligence sets out that providers of general-purpose AI models, such as LLMs, which are trained on large amounts of data, shall "draw up and make publicly available a sufficiently detailed summary about the content used for training of the general-purpose AI model."[210] Recital 107 explains that this summary "should be generally comprehensive in its

[206] Recital 6 of the Copyright Directive, *supra* note 117, expressly refers to the three-step step, stating that "[t]he exceptions and limitations provided for in this Directive seek to achieve a fair balance between the rights and interests of authors and other rightholders, on the one hand, and of users on the other. *They can be applied only in certain special cases that do not conflict with the normal exploitation of the works or other subject matter and do not unreasonably prejudice the legitimate interests of the rightholders*." (emphasis added).

[207] *See also* Senftleben, *supra* note 183, at 1545 (arguing that "[t]o ensure compliance with the three-step test, the potential adverse effect on the normal exploitation of a work can be minimized by giving right holders the opportunity to opt out").

[208] *See supra* Section III.A (discussing how copying copyrighted materials to form part of training datasets amounts to infringing acts of reproduction).

[209] This is also the case in the U.K. for its narrower text and data mining exception for non-commercial research purposes. *See* discussion *infra* Section III.B.3 (discussing Copyright, Designs and Patents Act 1988, c. 48, § 29A).

[210] EU AI Act, *supra* note 11, at art. 53(1)(d).

scope instead of technically detailed to facilitate parties with legitimate interests, including copyright holders, to exercise and enforce their rights under Union law, for example by 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." Because the summary is publicly available, rightholders should be able to identify if their works have been used as part of the training datasets, or if unclear, ask for further clarification. The AI Office, which is an administrative authoritative body set up to monitor compliance of the EU AI Act, should also monitor whether "the provider has fulfilled these obligations, but without verifying or proceeding to a work-by-work assessment of the training data in terms of copyright compliance."[211] It is presently unclear how that monitoring will be performed. Importantly, the obligation to comply with copyright law appears to be strict, regardless of model type or size, so long as it is classified as a general-purpose AI model.[212]

vi. Minimum Harmonization and Extended Collective Licensing

Although Articles 3 and 4 of the Copyright Directive are mandatory, they only set a minimum threshold for what exceptions and limitations may apply for text and data mining activities. Member States may therefore introduce more stringent rules if they so wish.[213] Member States are also in a position under Article 12 to provide for collecting licensing arrangements with extended effect. Extended collective licensing is a legal mechanism, which has a long history in Scandinavia, by which licensees enter into a single licensing agreement with a collective management organization. That licensing agreement is then extended to apply also to rightholders who have not authorized that collective management organization to represent them. Denmark recently decided to introduce a new bill for extended collective licensing for AI use of copyrighted works, including a meditation mechanism to facilitate negotiations between stakeholders.[214] Although the

[211] *Id.* at Recital 108.

[212] *Id.* at Recital 109 (stating that "*[w]ithout prejudice to Union copyright law*, compliance with these obligations should take due account of the size of the provider and allow simplified ways of compliance for SMEs, including start-ups, that should not represent an excessive cost and not discourage the use of such models." (emphasis added)).

[213] Copyright Directive, *supra* note 117, at art. 25.

[214] *See* Forslag til lov om ændring af lov om ophavsret og lov om videregående kunstneriske uddannelsesinstitutioner under Kulturministeriet, Bill No. L 145 (2023-1) (codified as amended at Lov om ophavsret [Danish Copyright Act], LBK nr. 1093 af 20. august 2023, § 50, stk. 2) (Den.); Danish Ministry of Culture, Bemærkninger til lovforslaget [Comments on the Bill], Kulturudvalget 2023-24, KUU Alm.del Exhibit 40, at 28 (Dec. 7 2023), https://www.ft.dk/samling/20231/almdel/KUU/bilag/40/2795790/index.htm (in English translation, "The proposed change means that in connection with the general extended collective license - e.g. with regard to collective agreements on text and data mining and artificial intelligence (AI) - mediation can be used to promote the conclusion of agreements. The parties themselves, i.e. the rights holders in specific areas of works and providers of AI services, can enter into and define agreements, and provided that the conditions are met, extended collective license effect can be attributed to them. This means that in the agreement that can be given extended collective license effect, it is possible to define which types of works, etc. are covered. . . . Exceptions in § 11b and § 11c do not provide full access to text and

mandatory text and data mining exceptions will still apply, extended collective licensing can effectively fill in the gaps for other works, such as works that have been opted-out or inadvertently infringing works forming part of the training datasets.

3. United Kingdom

The situation in the U.K. presents far greater difficulties for AI developers. U.K. law currently only allows text and data analysis for non-commercial research. Section 29A(1) of the Copyright, Designs and Patents Act 1988 (the "CDPA 1988") provides that copying for text and data analysis does not infringe copyright if there was lawful access to the work for the sole purpose of "research for a non-commercial purpose" and if the copy is accompanied by sufficient acknowledgement (unless this would be impossible for reasons of practicality or otherwise). Recital 42 of the Infosoc Directive, from which Section 29A originates, states that the non-commercial nature of an activity should be assessed based on the activity itself and that organizational structure and the means of funding of the establishment concerned are not the decisive factors in this respect. Commercial establishments could, therefore, in principle, conduct non-commercial research if it is done for a purpose not meant to generate revenue. However, in most circumstances, this is of little value to AI developers who generally have commercial interests in their products or services. AI developers in the U.K. also cannot avail themselves of copyright infringement by relying on another organization to collect the data sources on a non-commercial basis. Section 29A(2)(a) provides that copyright is infringed if the copy is transferred to any other person or used for any other purposes, unless authorized. This means that, if an AI developer has sourced datasets from a non-profit research organization, and the datasets contain copyrighted materials without authorization from rightholders, then the AI developer would still have committed copyright infringement.

The U.K. government launched a public consultation in 2021,[215] following which it announced that it intended to introduce a new data mining exception to support AI and such innovation activities in the U.K.[216] The suggested proposal did not have any possibility for rightholders to opt out of the provision and did not require developers to compensate rightholders for use of their content. The proposal was eventually withdrawn in 2023 after receiving significant backlash from

data mining of protected works. The general extended collective license [i.e., § 50, stk. 2 of the Danish Copyright Act] can be used here to fill the gaps by allowing collective agreements with extended license effect to the extent that use is not permitted under § 11b or § 11c.").

[215] *Open Consultation: Artificial Intelligence and Intellectual Property: copyright and patents*, U.K. INTELL. PROP. OFF. (Oct. 29, 2021), https://web.archive.org/web/20211101024732/https://www.gov.uk/government/consultations/artificial-intelligence-and-ip-copyright-and-patents/artificial-intelligence-and-intellectual-property-copyright-and-patents.

[216] *Consultation Outcome: Artificial Intelligence and Intellectual Property: copyright and patents: Government response to consultation*, U.K. INTELLECTUAL PROPERTY OFFICE (June 28, 2022), https://www.gov.uk/government/consultations/artificial-intelligence-and-ip-copyright-and-patents/outcome/artificial-intelligence-and-intellectual-property-copyright-and-patents-government-response-to-consultation [https://perma.cc/D28U-SPJF].

creative industries,[217] leaving AI developers with no legal recourse for avoiding copyright infringement at scale in the U.K. when training models.

4. Select Other Jurisdictions

There is substantial divergence among countries, other than the United States, EU, and the U.K., in the treatment of text and data mining for machine learning, including for generative AI models. Copyright exceptions and limitations are not harmonized to a significant extent. The Berne Convention only contains one mandatory provision concerning quotations[218] and a few optional provisions for other exceptions and limitations. This minimal harmonization has led to, as showcased above when comparing between the United States, EU, and the U.K., markedly different approaches in approaching text and data mining and copyright.

Most countries in the world have adopted exceptions and limitations that permit the reproduction of works for research purposes.[219] However, most countries have not adopted exceptions and limitations that expressly permit text and data mining, and only a very small number of countries have done so for non-research purposes,[220] like Singapore, which is discussed below. Although discussions concerning AI and copyright are actively pending at the World Intellectual Property Organization ("WIPO"), reaching international consensus on exceptions and limitations remains notoriously difficult due to different legal, cultural, social, and economic circumstances. Until then, if it happens at all, rightholders and developers will have to continue to face a patchwork of diverging copyright laws. The copyright laws of Japan, Singapore, Australia, India, and Israel are discussed below, as a few samples illustrating the differences in copyright law regarding text and data mining.

i. Japan

It is permissible in Japan to exploit copyrighted works in the course of computer data processing or other processes where the expressive portions of the work are

[217] CULTURE, MEDIA AND SPORT COMMITTEE, CONNECTED TECH: AI AND CREATIVE TECHNOLOGY: GOVERNMENT RESPONSE TO THE COMMITTEE'S ELEVENTH REPORT OF SESSION 2022-23, 2023-24, HC 441, at 4 (U.K.).

[218] Berne Convention, *supra* note 92, art. 10(1).

[219] *See* Michael Palmedo et al., *Measuring Change in Copyright Exceptions for Text and Data Mining* (Joint PIJIP/TLS Research Paper Series, Paper No. 98, 2023), https://digitalcommons.wcl.american.edu/cgi/viewcontent.cgi?article=1100&context=research [https://perma.cc/ZAX2-76YJ] (investigating the copyright statutes for 165 countries with respect to reproductions of works for research purposes).

[220] Raquel Xalabarder, *Scoping Study on the Practices and Challenges of Research Institutions and Research Purposes in Relation to Copyright*, AT 24, WIPO Doc. SCCR/44/4 (Oct. 17, 2023), https://www.wipo.int/edocs/mdocs/copyright/en/sccr_44/sccr_44_4.pdf (noting that"these TDM provisions are very unique and concentrate in a small number of jurisdictions. Most national laws do not formally exempt TDM uses, neither in general nor for research purposes").

not perceived by the human senses.[221] It is also permissible to exploit copyrighted works if it is done for use in data analysis, meaning the extraction, comparison, classification, or other statistical analysis of the constituent language, sounds, images, or other elemental data from a large number of works or a large volume of other such data.[222] Both of the two exceptions suppose, however, that the purpose of the exploitation is not to personally "enjoy" or cause another person to "enjoy" the thoughts or sentiments expressed in that work.[223] The exploitation must also not unreasonably prejudice the interests of the copyright owner in light of the nature or purpose of the work or the circumstances of its exploitation.[224] The term "enjoy" in this regard refers to obtaining the benefit of having the viewer's intellectual and emotional needs satisfied by using the copyrighted work.[225] How a work is enjoyed may differ depending on the context and the nature of the work. In a policy paper published by the Agency for Cultural Affairs of Japan in May 2024, it was considered that the reproduction of copyrighted works for AI training does not satisfy the "non-enjoyment purpose" requirement, and therefore infringes, where the "purpose of enjoyment" is *also* present. Such infringement risk could be present, it was stated, where the collection of works for AI training is used to generate materials that are similar to the sources used, such as fine-tuning, intentional "overfitting," or for the implementation of retrieval augmented generation ("RAG"). [226] Because copyright protects original expressions and not ideas, generated materials that merely resemble an author's style or look would not infringe, although this will always be a case-by-case analysis.[227] In sum, whether data reproduction for training generative AI models will infringe copyright will depend on the circumstances, in particular what is the intended output of the model and how similar outputs can become to the original source material through prompting.

ii. Singapore

The legal treatment in Singapore was initially similar to that in the United States —it had a flexible statutory fair use exception.[228] However, the fair use exception was considered unpredictable and unsatisfactory for the growing AI industry.[229]

[221] Chosakukenhō [Copyright Act], No. 121 of 2006, art. 30(4) (Japan), *translated in* (Japanese Law Translation [JLT DSJ]), https://www.japaneselawtranslation.go.jp/en/laws/view/1980/en [https://perma.cc/RG3C-SDEQ].

[222] *Id.* at art. 30(3).

[223] *Id.* at art. 30(1).

[224] *Id.*

[225] Japan Copyright Off., *supra* note 123, at 6.

[226] *Id.* at 8.

[227] *Id.* at 9.

[228] Copyright Act 2021, No. 22 of 2021, §§ 190-91 (Sing.).

[229] *See* Trina Ha et al., *When Code Creates: A Landscape Report on Issues at The Intersection of Artificial Intelligence and Intellectual Property Law*, INTELL. PROP. OFF. OF SINGAPORE 79 (Feb. 28, 2024), https://www.ipos.gov.sg/docs/default-source/resources-library/when-code-creates-landscape-report-on-ip-issues-in-ai.pdf [https://perma.cc/687V-FYFW].

Section 243 of the Singapore Copyright Act 2021 now provides an exception for computational data analysis, which includes using a computer program to identify, extract, and analyze information or data from a work or recording, and using the work or recording as an example of a type of information or data to improve the function of a computer program in relation to that type of information or data. This presupposes that the work or recording has been accessed through lawful means, meaning that, for example, paywalls have not been circumvented or the database's terms of conditions have not been breached.[230] The position in Singapore is therefore now similar to that of the EU but without any obligation to respect opt-outs from rightholders.

The fact that the Singaporean text and data mining exception does not have any provision that allows rightholders the possibility to reserve their rights might even be considered too favorable for developers and too unfavorable for rightholders. The three-step test in the Berne Convention and the TRIPS Agreement are meant to set the outer limits for how far copyright exceptions and limitations can go. A far-reaching text and data mining exception that makes it possible to scrape works publicly available online to create an AI model that *directly* competes with those same works could potentially "conflict with the normal exploitation of the work" and "unreasonably prejudice the legitimate interests of the author."

iii. Australia

There is no fair use exception or any specific text and data mining exception in Australia, making the situation difficult for AI developers. The only possible defenses for reproducing copyrighted materials as sources for training purposes relate to exceptions for fair dealing for research and study[231] or for temporary reproductions.[232] Both exceptions are narrow in scope and largely unhelpful to AI developers. Proposals to introduce text and data mining exceptions and limitations have yielded no results so far.

iv. India

There are similarly no specific exceptions or limitations for text and data mining activities in India. Instead, a statutory fair dealing exception applies, which is narrow in scope and limited to cases of private or personal use, including research, criticism, or review.[233] Sound recordings and cinematograph films also fall outside of the scope of fair dealing.[234] Although there have been discussions about introducing exceptions and limitations in the context of AI development, the latest position is to maintain the status quo and to require that permission is sought from rightholders for reproducing their works.

[230] Copyright Act 2021, No. 22 of 2021, § 244(2)(d) (Sing.).

[231] *Copyright Act 1968* (Cth) s 40 (Austl.).

[232] *Id.* at ss 43A, 43B.

[233] The Copyright Act, 1957, § 52(a) (India).

[234] Super Cassettes Indus. Vs. Chintamani Rao, 2011 SCC Online Del 4712 (India).

v. Israel

Like what is the case in the United States, Israel's Copyright Act includes a fair use exception.[235] The fair use test is worded slightly differently from the equivalent in the United States, however, and is reserved only for specifically permitted purposes, which include "private study" and "research" among others.[236] The Ministry of Justice of Israel published a non-binding opinion in December 2022 that discussed to what extent using copyrighted materials for machine learning was permitted under the fair use rule.[237] The opinion considered that machine learning is equivalent to the human process of inductive self-learning, such that it could count as a "private study" or "research" purpose.[238] The opinion then went on to consider each of the fair use factors. With respect to the transformative use of the works, it was considered that some machine learning systems are more transformative than others and that systems designed to produce outputs that highly resemble their inputs are unlikely to be transformative.[239] In other regards, however, it considered that machine learning datasets would often be transformative in nature given their societal value.[240]

The Ministry strictly interpreted the fourth fair use factor—the impact of the use on the value of the work and its potential market. In particular, it found that "[t]he expansion of machines' access to works will not harm existing markets of copyright owners, because such markets are nowhere to be found."[241] It also said that "[i]ndeed, even if ML enterprises intended to purchase licenses for each of the works in the dataset, doing so would be practically impossible. No platform offers such licenses, while the scope of the required works, their geographic distribution, and the lack of registration of rights in the works eliminate any possibility to obtain licenses directly from rightholders."[242] But even if such a license market would exist, the Ministry considered that "the price for licensing each work's would have reflected its marginal value in the dataset," which would be a very low, if any, profit, and which would not outweigh the "substantial economic and competitive gain" that could be generated from using the works in a machine learning dataset.[243]

[235] § 19, Copyright Act, 5768-2007, LSI 2199 34 (Isr.).

[236] *Id.* ("Fair use of a work is permitted for purposes such as: private study, research, criticism, review, journalistic reporting, quotation, or instruction and examination by an educational institution.").

[237] *Opinion: Uses of Copyrighted Materials for Machine Learning*, STATE OF ISRAEL, MINISTRY OF JUSTICE (Dec. 18, 2022), https://www.gov.il/BlobFolder/legalinfo/machine-learning/he/18-12-2022.pdf.

[238] *Id.* at 16-17.

[239] *Id.* at 18.

[240] *Id.* at 19.

[241] *Id.* at 20 (It is incorrect that there is no licensing market for training data. As previously mentioned, AI developers including OpenAI have signed licensing deals with selected data providers.) *See supra* Section III.B.1 (discussing the training data market in relation to the fourth fair use factor).

[242] *Opinion: Uses of Copyrighted Materials for Machine Learning*, *supra* note 237, at 20.

[243] *Id.* at 22.

Therefore, the Ministry concluded that, in most cases, the fair use doctrine would favor the use of copyrighted materials in a machine learning context.[244] However, as discussed above,[245] the position of what counts as fair use in the United States is arguably not as simple as the Ministry makes it appear to be in Israel, and there could be several situations where digital reproductions for model training is not fair use. The Ministry also fails to appreciate the collective economic value of using multiple, rather than individual, copyrighted works for training purposes.

C. Data Encoding and Memorization

1. Whether Generative AI Models "Memorize" Their Training Data

Machine algorithms function by learning from patterns and statistical correlations in the dataset.[246] Because of that, the output produced by generative AI models sometimes bears similarities to the original data that was used for training. It has also been shown that these models sometimes can, identically or near identically, reproduce the original data, in whole or in part. This raises the question of whether the original data itself is encoded and "memorized" in the model parameters, in which case, a second type of infringing act could arise. If the original data, which contains copyrighted materials, is somehow stored in the model parameters themselves, then even the model itself could be considered to have reproduced the works. If that is correct, then those separate acts of reproduction would require separate permissions from rightholders in order to not infringe their copyrights. Furthermore, if the generated output is identical or substantially similar to someone's works, then that is a further, separate copyright infringement.

Whether generative AI models store their training data, in whole or in part, and in any format and by any means, in their parameters or weights, is a hot topic in both courtrooms and data rooms. This is technically complex, because these models function almost like a black box, particularly LLMs. Even developers themselves do not presently have complete insight as to how these models work, such as which parameters are "activated" and how to generate a specific result.[247] This is a rapidly evolving research area. Only recently did Anthropic report a breakthrough, the so-called "Golden Gate Bridge" feature, where its researchers were able to link particular patterns of activities inside the LLM to both concrete and abstract concepts and by doing so were also able to control the behavior of the model.[248]

These technical uncertainties have not stopped litigants from making the relatively bold proposition that LLMs used to generate content retain the training data inside the model by memorization. For example, The New York Times stated

[244] *Id.*

[245] *See supra* Section III.B.1.

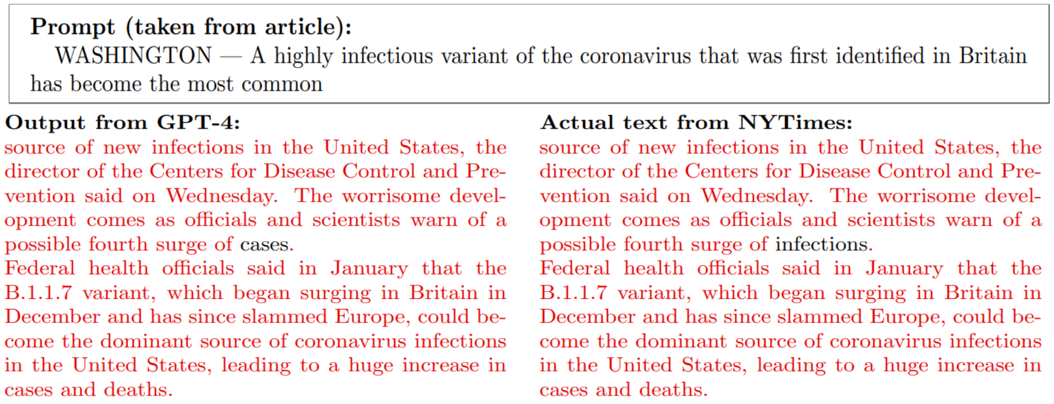
[246] *See supra* Section II.

[247] Vikas Hassija et al., *Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence*, 16 COGNITIVE COMPUTATION 45, 51 (2024).

[248] *Mapping the Mind of a Large Language Model*, ANTHROPIC (May 21, 2024), https://www.anthropic.com/research/mapping-mind-language-model [https://perma.cc/8V2T-TA4Q].

in its complaint against OpenAI in December 2023 that “the GPT LLMs themselves have ‘memorized’ copies of many of those same works encoded into their parameters” and that “the current GPT-4 LLM will output near-verbatim copies of significant portions of Times Works when prompted to do so.”[249] As evidence to support that, The New York Times submitted a compilation of 100 examples where ChatGPT-4 was able to reproduce large spans of text that were identical to the actual sources. Although this does not directly prove that the training data has been encoded inside the model’s parameters, it is indirect evidence suggesting the same. One such example is reproduced below in Figure 5, with verbatim copying shown in red.

Figure 5: Prompted excerpt from a New York Times article, and compared actual article text and output from ChatGPT-4[250]



Prompt (taken from article):
WASHINGTON — A highly infectious variant of the coronavirus that was first identified in Britain has become the most common

Output from GPT-4:
source of new infections in the United States, the director of the Centers for Disease Control and Prevention said on Wednesday. The worrisome development comes as officials and scientists warn of a possible fourth surge of cases.
Federal health officials said in January that the B.1.1.7 variant, which began surging in Britain in December and has since slammed Europe, could become the dominant source of coronavirus infections in the United States, leading to a huge increase in cases and deaths.

Actual text from NYTimes:
source of new infections in the United States, the director of the Centers for Disease Control and Prevention said on Wednesday. The worrisome development comes as officials and scientists warn of a possible fourth surge of infections.
Federal health officials said in January that the B.1.1.7 variant, which began surging in Britain in December and has since slammed Europe, could become the dominant source of coronavirus infections in the United States, leading to a huge increase in cases and deaths.

OpenAI responded by stating that unintended memorization, or regurgitation, occurs when the machine learning algorithm has not seen sufficient observations in the training data to enable a more generalized output.[251] OpenAI later also stated on its website that memorization was more common when particular content appears more than once in the training data.[252] The latter point is problematic because it means that more popular content is likely to be more prone to becoming memorized. Popular content is, of course, also typically highly valuable content from a copyright perspective.

[249] Complaint at 98, N.Y. Times v. Microsoft Corp., No. 23-cv-11195 (S.D.N.Y. filed Dec. 27, 2023).

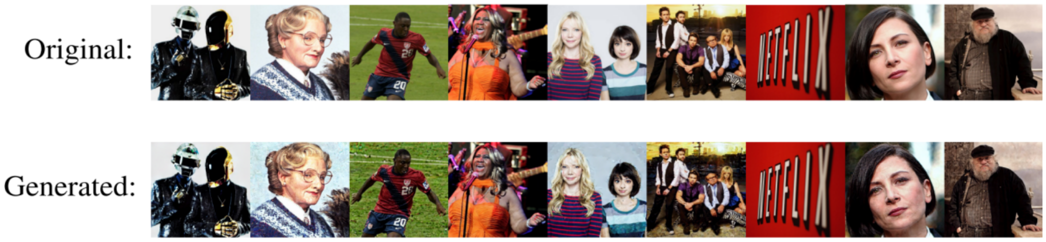
[250] Exhibit J, at 72-73, N.Y. Times v. Microsoft Corp., No. 23-cv-11195 (S.D.N.Y. filed Dec. 27, 2023).

[251] Memorandum of Law in Support of OpenAI Defendants’ Motion to Dismiss at 11, N.Y. Times v. Microsoft Corp., No. 23-cv-11195 (S.D.N.Y. filed Feb. 26, 2024) (referring to Gerrit J.J. van den Burg & Christopher K.I. Williams, *On Memorization in Probabilistic Deep Generative Models*, *in* PROCEEDINGS OF THE 35TH CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS 27916-27928 (Marc'Aurelio Ranzato eds., 2024)).

[252] *OpenAI and journalism*, *supra* note 151. *See also* Valentin Hartmann et al., *SoK: Memorization in General-Purpose Large Language Models*, ARXIV 9 (Oct. 24, 2023), https://arxiv.org/pdf/2310.18362 [https://perma.cc/Z6AT-7NV5] (noting that repetition of verbatim text in the training data makes LLMs more likely to memorize the sequence).

There is growing indirect evidence that not just text material may be memorized by LLMs. Images may also have been memorized from the training source material in diffusion-based models. Indeed, diffusion-based models, more than LLMs, are known to be prone to exhibiting memorization behavior.[253] This happens because the typical training objective of such models, which is to denoise score matching, has a closed-form optimal solution that is more prone to producing replicas of training data samples.[254] It has further been shown that caption duplication, which is the text used to describe a particular image, makes it more likely that the image will become memorized in the dataset.[255] In other words, if too many samples in the training dataset describe the same object in the same way, then the chance of duplication increases. One group of researchers found that, of 100,000 randomly sampled user-generated captions, approximately 1,200 had a similarity score above 0.5, which indicated that they may be duplicates.[256] A sample of generated versions of original images is shown below in Figure 6.

Figure 6: Images generated by Stable Diffusion v1.4 using random sampling and membership inference techniques, compared with originals[257]



2. When Encoding Training Data Will Infringe Copyright

A model is essentially a vast collection of parameters, which are encoded, numerical representations of the training data.[258] The parameters may represent the underlying content but cannot be understood as that content by humans without decoding. Yet there is no requirement in copyright law that copies must be immediately capable of human understanding. One way of looking at it is to compare it with the storage of digital copies. By analogy, an infringing copy will still have been made if a digital copy has been stored and encrypted on a secure

[253] Xiangming Gu et al., *On memorization in diffusion models*, ARXIV 1-2 (Oct. 4, 2023), https://arxiv.org/pdf/2310.02664 [https://perma.cc/JB3Q-6SJ3]; Gowthami Somepalli et al., *Understanding and Mitigating Copying in Diffusion Models,* ARXIV 1 (May 31, 2023), https://arxiv.org/pdf/2305.20086 [https://perma.cc/ET2E-PUDN].

[254] Gu et al., *supra* note 253, at 1.

[255] Somepalli et al., *supra* note 253, at 9.

[256] *Id.* at 2. For further studies conducted on the issue of memorization in AI models, which suggest that this is a rare technical occurrence except when fine-tuned datasets are employed, see references in *infra* notes 481-483.

[257] Carlini et al., *supra* note 57, at 5, fig. 3.

[258] *See supra* Section II. *See also* Cooper & Grimmelmann, *supra* note 7, at 33-38 (discussing how generative AI models can be deemed to contain encoded versions of its training data, sometimes literal copies).

server in a specific file format, even if that copy cannot be readily understood prior to altering the file format and deciphering the data.[259] This is because the right of reproduction extends to the reproduction of a work "in any manner or form."[260] The same argument can be stretched even further to say that even digital content that can readily be seen or heard simply represents a long binary sequence of 1s and 0s, the only difference being that we have a convenient medium to understand it.

But even if these arguments hold up in court, there is significant difficulty in producing evidence of what actually happens inside models. Although it is possible to generalize and broadly describe how a work, if part of the training data, would be used, we currently have no readily available means for investigating individual parameters and tracking those to particular training data. Courts are equipped to deal with evidentiary issues through presumptions of copying, but that appears unfair, as developers would need to undertake comprehensive research projects to disprove actual copying in the models.

More practically, there is a question of whether we should bother at all with asking whether reproductions are encoded into models themselves. If copyrighted materials have been copied in the course of model training, then infringement has already occurred. Finding another infringement would typically only result in nominal damages, or statutory damages where available, if it cannot be traced to any actual or potential loss.[261] While statutory damages are available as a recourse in the United States for copyright infringement, this is per infringing work, not per infringing act, and is elective instead of actual damages or profits.[262] On the one hand, it seems unlikely that there would be additional compensatory damages for reproductions in the model itself, either at all or at any meaningful amount. On the other hand, additional compensatory damages might be available where it can be shown that memorization is actually beneficial for the functioning of the model. Indeed, there are machine learning algorithms where training data is intentionally encoded as a feature, for example, *k*-nearest neighbor classification algorithms ("KNN") and support vector machines ("SVM").[263] Other such examples are

[259] Lee et al., *supra* note 7, at 74; Sobel, *supra* note 7, at 63-64.

[260] *See supra* Section III.A (referring to, *inter alia*, Article 9(1) of the Berne Convention, *supra* note 92).

[261] *See, e.g.*, Paramount Pictures Corporation v Hasluck [2006] 70 IPR 293 (Aust'l) (awarding only nominal damages for trademark infringement where there was no evidence of loss of reputation); Stoke-on-Trent City Council v. W & J Wass Ltd [1988] 1 WLR 1406, 1416 (U.K.) (generally in relation to breach of contract, holding that "[i]t is an established principle concerning the assessment of damages that a person who has wrongfully used another's property without causing the latter any pecuniary loss may still be liable to that other for more than nominal damages"); Wojnarowicz v. American Family Association, 745 F. Supp. 130, 149 (S.D.N.Y. 1990) (United States) (awarding nominal damages, where the plaintiff proved no actual damages for infringement of their moral rights); Toys "R" Us (Canada) Ltd v. Herbs "R" Us Wellness Soc'y, [2020] F.C. 682, ¶ 64 (Can. Ont.) (awarding nominal damages of $15,000 for trademark infringement where no evidence of actual monetary damage was filed, beyond the evidence of likely depreciation of goodwill).

[262] *See* 17 U.S.C. § 504(c)(1).

[263] Gavin Brown et al., *When is Memorization of Irrelevant Training Data Necessary for High-Accuracy Learning?*, ARXIV 3 (July 21, 2021), https://arxiv.org/pdf/2012.06421

intentional overfitting or RAG. RAG technologies, which dynamically pull content from external sources during generation,[264] inherently involve the use of potentially copyrighted material. Where memorization is an intended feature and not a technical bug, this should arguably qualify for additional compensatory damages. However, techniques such as RAG for text snippets also raise the question of whether copyright exceptions for quotation could apply, as these rely on extracting precise content from original sources. Such exceptions would require appropriate guardrails in place for length and context and that necessary acknowledgement is provided.

Secondary copyright infringement could also come into play where training data has been encoded into the model. The importation of infringing copies, while not harmonized in international copyright law, is an infringement under several copyright statutes, including in the United States,[265] U.K.,[266] and Singapore[267]. The argument that the completed model itself infringes copyright was recently raised by Getty Images in its pending complaint against Stability AI in the U.K.[268] In particular, this revolves around questions such as whether intangible information inside the Stable Diffusion model can be considered an "article" that can be "imported" and is "specifically designed or adapted for making [infringing] copies."[269] This is a highly fact-intensive argument, which may require showing, that each individual work has been encoded into the model parameters unless the burden of proof is reversed or loosened. This would be an extremely complex technical analysis for the reasons stated above and would be impossible to verify without developer access to the model.

The distinction between reproductions in the training datasets and in encoding into the model also becomes crucial where different parties engage in the different activities. If an AI model is fine-tuned for a domain-specific task, it is not always necessary to have access to the original dataset.[270] Access to the encoded parameters from the original dataset is then sufficient, which are then fine-tuned on

[https://perma.cc/8UUD-AER7]. KNN algorithms are commonly used for classification tasks or organizing existing data. *See* Ali Furkan Kalay, *Generating Synthetic Data with The Nearest Neighbors Algorithm*, ARXIV (Oct. 3, 2022), https://arxiv.org/pdf/2210.00884 [https://perma.cc/7BLK-YQ3Q]. SVM algorithms are commonly associated with classification and regression tasks and are similarly used to classify data. *See* Javier M. Moguerza & Alberto Munoz, *Support Vector Machines with Applications*, ARXIV 1 (Dec. 28, 2006), https://arxiv.org/pdf/math/0612817 [https://perma.cc/5YHX-BGBD].

[264] Yunfan Gao et al., *Retrieval-Augmented Generation for Large Language Models: A Survey*, ARXIV 1-3 (Mar. 27, 2024), https://arxiv.org/pdf/2312.10997 [https://perma.cc/6LSJ-A5T6].

[265] 17 U.S.C. § 602.

[266] Copyright, Designs and Patents Act 1988, c. 48, § 22 (U.K.).

[267] Copyright Act 2021, No. 22 of 2021, §147 (Sing.).

[268] Getty Images (US) Inc v. Stability AI Ltd [2023] EWHC (Ch) 3090 (U.K.). The issue of secondary copyright infringement for dataset creators and model providers is discussed in more detail in Section VIII.B, *infra*.

[269] Copyright, Designs and Patents Act 1988, §24(1) (U.K.). This is also discussed in *infra* Section VIII.B.

[270] *See supra* Section II.

supplemental and specialized datasets. This could become a critical issue where open-source LLMs, such as ChatGPT, Claude, or Mistral AI, are employed by other parties to create their own custom models. If training data is encoded into the model parameters themselves, then those third parties could also become liable for massive copyright infringement and separate damages. When developers fine-tune a pre-trained LLM, they begin by downloading the pre-trained model's weights, which are the parameters of the model.[271] Therefore, although custom developers do not necessarily reproduce the original training data itself, they might reproduce encoded versions of the training data as a necessary activity as part of the model fine-tuning process.[272] If this is correct, then not only AI developers but hundreds of thousands, if not millions, of custom developers could potentially have committed copyright infringement, assuming the "memorization" argument actually holds up in court. Therefore, the significance of this discussion cannot be understated.

Last but not least, there is another, more precarious argument that could be raised about the extent to which AI models "memorize" their training data in an infringing way. If the training data is encoded into the model parameters, and if the model is distributed to the public, then there is an argument to be made that the infringing training data, in encoded format, is "communicated" to the public.[273]

[271] *See*, e.g., *Fine-Tune Your First LLM*, PYTORCH https://pytorch.org/torchtune/stable/tutorials/first_finetune_tutorial.html [perma.cc/4HVS-P6DM] (last visited Oct. 26, 2024) (describing how to fine-tune an LLM, and stating that it is first necessary to "be granted access in order to download the weights"); Venkatesh Balavadhani Parthasarathy et al., *The Ultimate Guide to Fine-Tuning LLMs from Basics to Breakthroughs: An Exhaustive Review of Technologies, Research, Best Practices, Applied Research Challenges and Opportunities*, ARXIV 22 (Oct. 30, 2024), https://arxiv.org/pdf/2408.13296 [https://perma.cc/2W26-WPYX] (stating that it is necessary, in order to fine-tune a model, to "[u]se the chosen framework's functions to download the pre-trained model from an online repository"); Jayesh Suthar, *The Ultimate Guide to Fine-Tuning Large Language Models with Hugging Face*, MEDIUM (May 27, 2024), https://medium.com/@jayeshchouhan826/the-ultimate-guide-to-fine-tuning-large-language-models-with-hugging-face-c971e588bf02 [perma.cc/8RZT-8LJT] (stating that, in a tutorial for fine-tuning LLMs, "we will be using HuggingFace libraries to download and train the model. To download models from HuggingFace, we will need an Access Token.").

[272] Keita Kurita et al., *Weight Poisoning Attacks on Pre-trained Model*, *in* PROCEEDINGS OF THE 58TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS 2793, 2793 (Dan Jurafsky et al. eds., 2020) (stating that "[t]raining these large models is computationally prohibitive, and thus practitioners generally resort to downloading pre-trained weights"). *Models and pre-trained weights*, PYTORCH, https://pytorch.org/vision/stable/models.html [perma.cc/DWT9-FCR9] (last visted Oct. 26, 2024) (stating that "[i]nstancing a pre-trained model will download its weights to a cache directory"). For example, to download Stable Diffusion v1.4's pre-trained weights, see stable-diffusion-v-1-4-original, HUGGING FACE, https://huggingface.co/CompVis/stable-diffusion-v-1-4-original [perma.cc/5CBQ-5RQC] (last visited Oct. 26, 2024). *See also* Cooper & Grimmelmann, *supra* note 7, at 33-38 (discussing how generative AI models, when encoding "patterns" of its training data, can sometimes encode literal copies of the data).

[273] *See*, e.g., Infosoc Directive, *supra* note 97, at art. 3(1); 17 U.S.C. § 106(3) (United States); The Copyright, Designs and Patents Act 1988, c. 48 § 20(1). § 20(2) (U.K.) also expressly provides that the communication to the public includes "the making available to the public of the work by electronic transmission in such a way that members of the public may access it from a place and at

This will fail or win depending on the legal interpretation of when works are deemed "communicated" and if "works" should be technically neutral, extending to copies of works in any form, in whole or in part, in the same way as the right of reproduction.

3. Exceptions and Limitations That Apply to Data Encoding

As we have already seen above for data reproductions in the course of model training, there is substantial divergence between jurisdictions regarding copyright exceptions and limitations. That divergence is mirrored for data encoding-related infringements. The same exceptions and limitations, or the absence of such, that apply for text and data mining activities will also apply for data encoding in models. However, whether that means that the outcome is the same is another question.

In the United States, it is unlikely that the fair use test would have a different outcome where the allegedly infringing party, the developer or dataset creator, committed both infringing acts. Although the encoded parameters are the result of a further technical step taken when developing the model, their use may be considered minimally transformative when encoded parameters are a replica of training data, the difference being just the different formats.[274] Similarly, the fourth fair use factor of assessing market impact is less likely to be met when training data has been memorized verbatim; it is subsequently capable of being generated verbatim and targets the same audience as the original author. The generated output of such encoded, memorized data will then directly compete with the original works.[275]

Where the defendant is the intermediate user of the model; however, different considerations arise in the fair use test. For example, a custom developer who downloads model parameters to fine-tune them and build a custom model, may not necessarily use the model to take advantage of any memorized copies of works. If the custom developer uses the model parameters for purposes other than to reproduce encoded, memorized data that harms or potentially harms original

a time individually chosen by them. *See also* Lee et al., *supra* note 7, at 71 (arguing that the distribution right could be implicated when a model trainer makes the model available for download).

[274] This was considered by the Ministry of Justice of Israel in its analysis of its equivalent fair use test. *See Opinion: Uses of Copyrighted Materials for Machine Learning*, *supra* note 237, at 18 ("In particular, systems can be designed to produce outputs that would highly resemble their inputs. This can be done deliberately (for example, when a system is designed to imitate a specific author or a genre), or unintentionally, such as when the dataset is not sufficiently diverse. In such circumstances, the use may not be considered transformative."). *See also* Lee et al., *supra* note 7, at 106; Jacob Alhadeff et al., *Limits of Algorithmic Fair Use*, 19 WASH. J. L. TECH. & ARTS. 1, 43-44 (2024) (arguing that a verbatim or near-verbatim generated copy of the input, which targets the same audience, should require a more significant change in the content to be considered transformative).

[275] The fourth fair use factor requires the court to consider "not only the extent of market harm caused by the particular actions of the alleged infringer, but also 'whether unrestricted and widespread conduct of the sort engaged in by the defendant . . . would result in a substantially adverse impact on the potential market' for the original." *See* Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 590 (1994).

authors, then there will be a stronger argument that their reproduction of the works in encoded format is fair use.

In the EU, text and data mining is broadly defined in Article 2(2) of the Copyright Directive as any "automated analytical technique" aimed at analyzing text and data in digital form "in order to generate information," which includes but is not limited to patterns, trends and correlations. It is likely that data encoding, when executed by the developer who uses the original dataset to create the pre-trained model parameters, would also be considered an automated analytical technique. This process is "automated" as it is performed by computers instead of humans,[276] and "analytical" because encoding data involves computationally processing original data.[277] Encoding data into model parameters is also done for the purpose of generating new information in the context of machine learning algorithms.[278] Indeed, encoding training data into model parameters that are statistically meaningful is a necessary processing step for many machine learning algorithms.[279]

This means that developers who are reproducing or extracting copyrighted materials in encoded format will likely be able to benefit from the text and data mining exception in Article 4(1) of the Copyright Directive. The same will also be true for custom developers who fine-tune pre-trained models and who download the model parameters, even if the models contain reproductions of the training data in encoded format. In principle, this also means that custom developers will need to respect any opt-outs from rightholders.[280] However, this will become extremely difficult, if not impossible, as custom developers have no easy way of knowing what original training data was used by the original developers who pre-trained the

[276] *Cf.* Triaille et al., *supra* note 166, at 17 (defining "automated" in the context of text and data mining as data analysis that is done using computers).

[277] *Id.* (defining "data analysis" as the processing of data, which may include the extraction, copy, comparison, classification or other statistical analysis, etc. of data, or a mix of them).

[278] *Id.* at 17-18 (defining "to uncover new knowledge or insights" as uncovering new knowledge, new insights, new relationships, etc.). *See also supra* Section II (discussing, with references, the purpose of machine learning algorithms for learning patterns and making predictions).

[279] Liu et al., *supra* note 16, at 10 (discussing that "[an LLM] model observes large amounts of textual data and attempts to predict the next word at each position in the text. This gradual learning process allows the model to capture the patterns and information inherent in language, encoding a vast amount of linguistic knowledge into its parameters.").

[280] One exception is Recital 109 of the EU AI Act, *supra* note 11, which could be read as excluding the obligations for providers of fine-tuned or modified models to only the new training dataset, and not the original dataset used for the general-purpose AI model ("[i]n the case of a modification or fine-tuning of a model, the obligations for providers of general-purpose AI models should be limited to that modification or fine-tuning, for example by complementing the already existing technical documentation with information on the modifications, including new training data sources, as a means to comply with the value chain obligations provided in this Regulation"). However, it is not clear whether those excluded obligations also extend to copyright law.

model. This puts custom developers at the risk of infringing copyright without having any means of avoiding it.[281]

IV. TERRITORIALITY OF COPYRIGHT LAW

There is no such thing as international copyright law. There is no single international statute or any unitary right that governs copyright protection in all countries. What does exist is a form of minimum harmonization through international treaties, including the Berne Convention and the TRIPS Agreement. These conventions set out a minimum threshold for what types of works should be protected, what exclusive rights should be granted, and provide a limited number of exceptions and limitations, most of which are optional for signatory states to adopt.[282] The minimum level of harmonization has led to national copyright statutes that, while sharing similar structures and features, differ significantly in their details. Therefore, authors enjoy a bundle of distinct national rights that automatically arise in each country of protection from the moment of creation.[283] This fundamental principle of national interdependence is confirmed in Article 5 of the Berne Convention. Specifically, Article 5(2) provides that authors shall enjoy and exercise their rights in all signatory states, independently of the existence of protection in the country of origin of the work.

The fact that copyright is territorial further means, as also stated in Article 5(2) of the Berne Convention, that the extent of protection is governed exclusively by the laws of the country where protection is claimed. This long-standing rule has been commonly interpreted as referring to *lex loci protectionis*[284] and derives from the principle of territoriality. Protection is claimed in the jurisdiction where the infringing activity is deemed to have occurred, which is a question of substantive law that is to be interpreted in each national copyright regime. In practice, *lex loci protectionis* means that the country where the infringing activity has occurred shall regulate the infringing behavior, unless that same activity can be localized in another country as well.[285] Simply put, infringing activities shall only be governed

[281] Practically, this can be resolved by introducing warranties and indemnities into license agreements for the use of pre-trained models. Original developers would then become contractually liable towards custom developers for any infringing model parameters.

[282] *See, e.g.*, Graeme Dinwoodie, *The Development and Incorporation of International Norms in the Formation of Copyright Law*, 62 OHIO ST. L.J. 733, 737-40 (2001) (describing the international copyright system as one based on minimum standards of copyright protection, together with territorial rights for which countries owe a duty of national treatment towards foreign authors).

[283] *See, e.g.*, Jane C. Ginsburg, *International Copyright: From a "Bundle" of National Copyright Laws to a Supranational Code?*, 47 J. COPYRIGHT SOC'Y U.S.A. 265, 266 (2000).

[284] *See* SAM RICKETSON & JANE GINSBURG, INTERNATIONAL COPYRIGHT AND NEIGHBOURING RIGHTS: THE BERNE CONVENTION AND BEYOND §§ 20.02-20.03 (3d. ed., 2022); Alexander Peukert, *Territoriality and Extraterritoriality in Intellectual Property Law*, *in* TRANSNATIONAL LEGAL AUTHORITY IN AN AGE OF GLOBALIZATION 189, 191-192 (Günther Handl ed., 2012); Raquel Xalabarder, *Copyright: Choice of Law and Jurisdiction in the Digital Age*, 8 ANN. SURV. INT'L & COMP. L. 79, 82-83 (2002).

[285] Mattias Rättzén, *Closing the Patent Loophole Across Borders*, 20 UIC REV. INTELL. PROP. L. 358, 370 (2021).

by the copyright laws of the country where the infringement is deemed to have taken place. This means that domestic copyright laws only apply within the boundaries of their respective country of protection.[286] Indeed, as the European Court of Justice stated in its landmark *Lagardère* decision, "it is clear from its wording and scheme that Directive 92/100 provides for minimal harmonization regarding rights related to copyright. Thus, it does not purport to detract, in particular, from the principle of the territoriality of those rights, which is recognized in international law and also in the EC Treaty. Those rights are therefore of a territorial nature and, moreover, domestic law can only penalize conduct engaged in within national territory."[287]

Since copyright exists separately and independently in each country of protection, as a legal construction dictated by national law, by definition, it cannot afford protection anywhere else.[288] However, it would be a misunderstanding to assume that copyright law is therefore strictly territorially-limited. It is commonly recognized that the principle of territoriality does not in itself bar taking facts that have occurred abroad but have legal significance into account.[289] A cross-border infringement is nothing more than a distribution of such facts spanning over multiple jurisdictions.[290] The relevant question then becomes whether those activities are sufficient to conclude that an infringement has occurred in a particular country of protection, also known as localization.[291] That is to say, the only leeway that regulators and courts have is to define or interpret more broadly, or narrowly, the circumstances under which infringement is deemed to have occurred within

[286] Xalabarder, *supra* note 284, at 83 ("Domestic copyright laws apply within the boundaries of each respective state. There exist as many copyrights as copyright domestic laws."); Jane C. Ginsburg & Morton L. Janklow, *Private International Law Aspects of the Protection of Works and Objects of Related Rights Transmitted through Digital Networks*, at 34, WIPO Doc. GCPIC/2 (Nov. 30, 1998) (stating that "to apply anything less than the laws of all the countries where the alleged infringement occurred would be to disregard the basic nature of international copyright under the Berne Convention.").

[287] Case C-192/04, Lagardère Active Broadcast, ECLI:EU:C:2005:475, ¶ 46 (July 14, 2005).

[288] *See, e.g.*, Peukert, *supra* note 284, at 189; Eugen Ulmer, *General Questions–the International Conventions*, 14 INTERNATIONAL ENCYCLOPEDIA OF COMPARATIVE LAW: COPYRIGHT 5 (Eugen Ulmer & Gerhard Schricker ed., 2007); Jürgen Basedow, *Foundations of Private International Law in Intellectual Property, in* INTELLECTUAL PROPERTY IN THE GLOBAL ARENA: JURISDICTION, APPLICABLE LAW, AND THE RECOGNITION OF JUDGMENTS IN EUROPE, JAPAN AND THE US 3, 7-8 (Jürgen Basedow et al., ed., 2010); Marketa Trimble, *Advancing National Intellectual Property Policies in a Transnational Context*, 74 MD. L. REV. 203, 231 (2015).

[289] *See, e.g.*, Rättzén, *supra* note 285, at 371; Rättzén, *supra* note 201, at 544-45; Peukert, *supra* note 284, at 198-202; Alexander von Muhlendahl & Dieter Stauder, *Territorial Intellectual Property Rights in a Global Economy–Transit and Other "Free Zones"*, *in* PATENTS AND TECHNOLOGICAL PROGRESS IN A GLOBALIZED WORLD: LIBER AMICORUM JOSEPH STRAUS 654 (Prinz zu Waldeck und Pyrmont et al. eds., 2009); Roberto Romandini & Alexander Klicznik, *The Territoriality Principle and Transnational Use of Patented Inventions–The Wider Reach of a Unitary Patent and the Role of the CJEU*, 44 INT'L REV. INTELL. PROP. & COMPETITION L. 524, 530 (2013); Peter Mankowski, BRUSSELS I REGULATION: EUROPEAN COMMENTARIES ON PRIVATE INTERNATIONAL LAW 201 (Ulrich Magnus & Peter Mankowski eds., 1st ed. 2007).

[290] Rättzén, *supra* note 285, at 369-70.

[291] *Id.* at 370.

their borders. Localizing infringing activities in cross-border situations is not an optional recourse.[292] Indeed, *lex loci protectionis* as a choice of law rule makes localization of the infringement necessary to the choice of law analysis.[293] If protection is claimed in one country but infringement occurs in another, no infringement will deemed to have occurred in the country whose laws are supposed to apply, resulting in the subsequent dismissal of the case on the merits. The country where infringement is asserted to have occurred and the country whose laws are applicable must therefore coincide, if we are adhering to *lex loci protectionis* as the choice of law rule.[294]

The localization of cross-border copyright infringement is not harmonized in any international conventions, and national copyright statutes are often silent on this issue. It is instead a topic typically reserved for courts to resolve. Unsurprisingly, this has led to vast fora with different approaches in different jurisdictions, where some courts are more or less inclined to assume prescriptive jurisdiction over foreign infringing conduct.[295] This is not just a legal analysis but also a contextual analysis, which will depend on what type of infringing conduct is concerned. Broadly speaking, courts have adjudicated infringing conduct in two types of scenarios: where an infringing act was commenced within the territory but consummated abroad (also known as subjective territoriality) and where an infringing act was commenced abroad but consummated within the territory (also known as objective territoriality). Those two bases for asserting jurisdiction are not exclusive to copyright law, let alone intellectual property law, but apply to all laws.[296] Sovereignty is a universal concept in public international law. It not only grants states the right to legislate within their own territories, subject to finding a sufficient territorial connection[297] but also a duty to recognize the right of other

[292] *Id.*

[293] *See, e.g.*, Rättzén, *supra* note 201, at 543; Andreas P. Reindl, *Choosing Law in Cyberspace: Copyright Conflicts on Global Networks*, 19 MICH. J. INT'L L. 799, 803-4 (1998); Nathan R. Wollman, *Maneuvering Through the Landmines of Multiterritorial Copyright Litigation: How to Avoid the Presumption Against Extraterritoriality When Attempting to Recover for the Foreign Exploitation of U.S. Copyrighted Works*, 104 W. VA. L. REV. 343, 362 (2002); Christian M. Rieder, *U. S. Subject Matter Jurisdiction for Copyright Infringements on the Internet*, COMPUT. L. REV. & TECH. J. 23, 28 (1998); Xalabarder, *supra* note 284, at 82-83.

[294] Rättzén, *supra* note 285, at 370.

[295] *Id.* at 372-73 (summarizing the different approaches to localizing intellectual property infringements taken by different courts); *see also* Lydia Lundstedt, *Territoriality in Intellectual Property Law* 434-62 (Oct. 4, 2016) (Ph.D. dissertation, Stockholm University) (discussing different approaches taken by different courts in the EU, U.K. and the United States in the context of copyright inbound and outbound regulation).

[296] Rättzén, *supra* note 285, at 362-363 (referring to the landmark case from the Permanent Court of International Justice, S.S. Lotus (Fr. v. Turk.), Judgment, 1927 P.C.I.J. (ser. A) No. 10 (Sept. 7)). *See also* Harold G. Maier, *Jurisdictional Rules in Customary International Law*, *in* EXTRATERRITORIAL JURISDICTION IN THEORY AND PRACTICE 64, 65-69, 83-84, 90 (Karl M. Meessen ed., 1996); Cedric Ryngaert, *Jurisdiction in International Law* 38 (Feb. 12, 2007) (Ph.D. dissertation, Leuven University) (discussing more generally the public international law basis for asserting legislative jurisdiction).

[297] Rättzén, *supra* note 285, at 362-363, 370; S.S. Lotus*,* 1927 P.C.I.J. (ser. A) No. 10, at 18-19 (holding that "jurisdiction is certainly territorial; it cannot be exercised by a State outside its

states to legislate within their respective territories.[298] As phrased by Justice Joseph Story of the U.S. Supreme Court in 1824, and still as relevant today, "[t]he laws of no nation can justly extend beyond its own territories, except so far as regards its own citizens. They can have no force to control the sovereignty or rights of any other nation, within its own jurisdiction."[299]

Sovereignty is not a one-sided concept. Indeed, as Professor Graeme Dinwoodie has phrased it, "the purest act of sovereignty is to foreswear from acting in circumstances in which it is descriptively and prescriptively possible to do so."[300] It would be a gross, even incorrect, simplification to say that prescriptive jurisdiction is merely about identifying connecting factors to a country's territory. Even if there is a sufficient nexus constituting a jurisdictional basis, the exercise of that prescriptive jurisdiction must not interfere with that of another state.[301] Simply put, it should be queried whether prescribing foreign behavior would amount to a legal conflict or result in adverse offshore private or public interests.[302] Of course, this is easier said than done, and courts unfortunately typically never proceed to this step when assuming prescriptive jurisdiction. In most cases, courts stop the analysis once they have established a sufficient connecting factor to their own forum.[303]

territory except by virtue of a permissive rule derived from international custom or from a convention"); Frederick A. Mann, *The Doctrine of Jurisdiction in International Law*, 111 RECUEIL DES COURS 1, 46 (1964) (describing it as, "[s]ince the doctrine of international jurisdiction is not at present concerned with exclusivity of jurisdiction, the legally relevant point of contact will have to be defined as indicating the State which has a close, rather than the closest, connection with the facts, a genuine link, a sufficiently strong" and the relevant question as whether "the contact is sufficiently close"). Mann's description of the concept of jurisdiction in international law is described as the "classical doctrine" of international jurisdiction. Ryngaert, *supra* note 296, at 133.

[298] Rättzén, *supra* note 285, at 370; *see also* Island of Palmas (U.S. v. Neths.), 2 R.I.A.A. 829, 838-39 (Perm. Ct. Arb. 1928).

[299] *The Appollon*, 22 U.S. (9 Wheat.) 362, 370 (1824).

[300] Graeme B. Dinwoodie, *Developing a Private International Intellectual Property Law: The Demise of Territoriality?*, 51 WM. & MARY L. REV. 711, 773 (2009).

[301] Island of Palmas, 2 R.I.A.A. at 838-39 (holding that territorial sovereignty has "as corollary a duty: the obligation to protect within the territory the rights of other States"); Barcelona Traction, Light & Power Co. (Belg. v. Spain), Judgment, 1970I.C.J. REP. 3, =105, ¶ 70 (February 5) (separate opinion by Fitzmaurice, J.) (holding that state jurisdiction is not unlimited in public international law, as it involves "for every State an obligation to exercise moderation and restraint as to the extent of the jurisdiction assumed by its courts in cases having a foreign element, and to avoid undue encroachment on a jurisdiction more properly appertaining to, or more appropriately exercisable by, another State."); *see also* Ryngaert, *supra* note 296, at 40; Mann, *supra* note 297, at 31-32.

[302] For a more detailed analysis (generally, and in the context of patent law) of how to balance domestic and foreign sovereignty, including both competing territorial connections and competing interests, *see* Rättzén, *supra* note 285, at 378-79, 404-10.

[303] American courts are a noteworthy exception, which more frequently than others have referred to foreign sovereignty and public interests as a limiting factor when assuming prescriptive jurisdiction. For example, a relative impact analysis is part of the *Timberlane* factors considered in antitrust and trademark law. *See* Reebok Int'l Ltd. v. Marnatech Enters., Inc., 970 F.2d 552, 555 (9th Cir. 1992); Timberlane Lumber Co. v. Bank of Am. Nat'l Trust & Sav. Ass'n, 549 F.2d 597, 614 (9th Cir. 1976) (both cases explaining that the elements to be weighed include, amongst others, "the degree of conflict with foreign law or policy, . . . relative significance of effects on the United

The right of reproduction in copyright law has not caused much difficulty for courts to localize. It is typically clear from both a legal and factual point of view that an act of reproduction has occurred, and most such acts will take place in one country only. For example, if a protected work is downloaded from a server or duplicated on a server in country A, then the act of reproduction will have occurred in country A. More dubious cases arise where the person downloading or uploading content on a server is located in country A but the server is located in country B. Where there are multiple connecting factors in a given case, then multiple states may have the right to concurrently regulate that conduct independent of each other[304] as a natural consequence of each state's right to apply their laws to conduct that, in whole or in part, occurs within its territory. This is often the case for online conduct and cross-border copyright infringement. Indeed, the Canadian Supreme Court has held that: "[i]n terms of the Internet, relevant connecting factors would include the situs of the content provider, the host server, the intermediaries and the end user . . . . [T]hat Canada could exercise copyright jurisdiction in respect [to] both of transmissions originating here, and transmissions originating abroad but received here, is not only consistent with our general law but with both national and international copyright practice."[305] Similar views have also been favored by American and European courts, attaching significance to each connecting factor relevant for the infringing conduct.[306] In *Lagardère*,[307] referred to above, a French broadcasting company had transmitted a signal from France, aimed at a French audience, but using a transmitter in Germany to enhance its reach within France. Although the signals could also be received in a limited area in Germany, there was no commercial exploitation of the broadcasts in Germany. In that case, the European Court of Justice famously ruled that neither EU law nor international law

States as compared with those elsewhere, . . . and the relative importance to the violations charged of conduct within the United States as compared with conduct abroad."). The reason for this unique trend in American case law originates from the fact that the U.S. Supreme Court has repeatedly held that in the absence of a contrary, clear and affirmative intent suggesting otherwise, United States legislation only operates within the territorial jurisdiction of the United States. *See* EEOC v. Arabian American Oil Co., 499 U.S. 244, 248 (1991). *See also* Morrison v. Nat'l Austl. Bank Ltd., 561 U.S. 247, 255, 261 (2010).

[304] Laker Airways Ltd. v. Sabena, Belgian World Airlines, 731 F.2d 909, 952 (D.C. Cir. 1984) (stating that "[t]here is no principle of international law which abolishes concurrent jurisdiction"). *See also* Ryngaert, *supra* note 296, at 135.

[305] Soc'y of Composers, Authors & Music Publishers of Can. v. Canadian Ass'n. of Internet Providers, 2004 SCC 45, paras. 61, 76 (Can.).

[306] *See* Nat'l Football League v. PrimeTime 24 Joint Venture, 211 F.3d 10, 13 (2d Cir. 2000) (holding that a public performance or display includes "each step in the process by which a protected work wends its way to its audience," citing *David v. Showtime/The Movie Channel, Inc.*, 697 F. Supp. 752, 759 (S.D.N.Y. 1988)); Case C-5/11, Titus Alexander Jochen Donner, ECLI:EU:C:2012:370, ¶¶ 26-27 (June 21, 2012) (holding that the distribution right in copyright is characterized by a series of acts going "at the very least" from the conclusion of a contract of sale to the performance thereof by delivery to a member of the public, and that acts giving rise to a distribution to the public may therefore take place in a number of member states); EMI Records Ltd. v British Sky Broadcasting Ltd. [2013] EWHC (Ch) 379 [35]-[38] (the act of communication to the public can be localized to the U.K. where the uploading party is located in the U.K., notwithstanding that the server is located abroad).

[307] Case C-192/04, Lagardère Active Broadcast, 2005 E.C.R. I-7218, ¶ 46.

prevents Member States from localizing the act of broadcasting within their respective territories if transmitters located in those states are used, even if the broadcast has minimal or no impact on the rightholder's exclusive market in those states.[308]

But courts have not only looked at connecting factors and how proximate they are when localizing intellectual property infringements. Maintaining the efficacy of intellectual property laws in cross-border contexts has also been considered.[309] Courts have long been attentive to the real possibility of avoiding intellectual property laws by arbitrarily relocating in whole or in part the infringing conduct to another country. This has in turn reinforced the motivation for courts to look at a multitude of potential connecting factors when assessing whether the infringing conduct can be localized domestically. For example, the European Court of Justice held in *L'Oréal SA v. eBay* that there "must" be trademark infringement when trademark protected goods were offered for sale from a third country to consumers in the country of protection.[310] Otherwise, online operators could simply relocate their establishment or servers abroad while still accessing the home market, which would "undermine" the effectiveness of trademark laws.[311] Market repercussions are also relevant, in particular whether offshore activities still produce an effect in the forum in question. In the patents context, some courts have considered if the alleged infringer can still reap the commercial benefits of the infringing activity in the forum, even if activities are occurring abroad.[312] Similarly, courts have long

[308] *Id.* ¶¶ 46, 53-55.

[309] *See* Rättzén, *supra* note 285, at 383 (discussing efficacy as theme when localizing infringements in the context of patent law).

[310] Case C-324/09, L'Oréal SA v. eBay, 2011 E.C.R. I-6073, ¶¶ 61-62.

[311] *Id.* at ¶¶ 62-63. A similar explanation was provided in Case C-173/11, Football Dataco v. Sportradar, ECLI:EU:C:2012:642, ¶¶ 44-47 (October 18, 2012) (rejecting the argument that an act of re-utilisation must be located exclusively to the territory of the member state where the web server is located from which the data in question is sent, which would impair the effectiveness of the protection afforded under that national law).

[312] *Prepaid-Karten II*, OLG, Dec. 10, 2009, 2 U 51/08, openJur (Ger.) https://openjur.de/u/143082.html [https://perma.cc/BQ8F-ARQH] (finding direct patent infringement in Germany of a method claim, where the relevant infringing acts were performed on foreign servers, but where the foreign acts were deemed to be intended to have an effect, in the economic sense, in Germany); Decca, Ltd. v. United States, 544 F.2d 1070, 1083 (Ct. Cl. 1976) (finding that a patent concerning a radio navigation system was infringed even if one of the transmitting stations, as part of the claim, was located abroad, since the equipment abroad was still owned and "controlled" by the defendant, and since the "actual beneficial use" of the system was domestic).

looked at market effects in the trademarks context,[313] as well as in copyright law,[314] when localizing cross-border infringements. This has particularly been the case when foreign conduct produces an effect on domestic commerce,[315] which is indicative of harm to the rightholder in that particular country. This is a form of inbound regulation, where rightholders are protected from spillover effects resulting from conduct originating from abroad.[316]

The best example of regulating foreign behavior that has an inbound impact in copyright law and trademark law is the targeting doctrine. Copyright law grants authors a right to communicate their works to the public.[317] If a work is communicated on the internet, which is a ubiquitous environment that can be accessed anywhere in the world, then this gives rise to the possibility of there being a great number of "publics," in each country which are being communicated to. If mere accessibility was sufficient to assume prescriptive jurisdiction in this regard, then those who upload works on the internet could potentially become liable for copyright infringement everywhere in the world at the same time.[318] This has prompted courts to narrow what territorial connecting factors are deemed sufficient for cross-border online infringements in an effort to avoid potentially massive

[313] *Hotel Maritime*, Bundesgerichtshof [BGH] [Federal Court of Justice] Oct. 13, 2004 Gewerblicher Rechtsschutz und Urheberrecht, Internationaler Teil [GRUR Int.] 431 (2005) (Ger.) (recognizing that the commercial effects of the use of a foreign trademark in Germany were too weak to find for trademark infringement); Steele v. Bulova Watch Co., 344 U.S. 280, 285-287 (1952); Vanity Fair Mills, Inc. v. T. Eaton Co., 234 F.2d 633, 641-643 (2d Cir. 1956) (both cases finding that trademark infringement within the United States supposes that there is a substantial effect on commerce within the United States); Case C-324/09, L'Oréal SA v. eBay, 2011 E.C.R. I-6073, ¶¶ 61-67 (July 12, 2011) (trademark infringement within the EU on the internet supposes that consumers within the EU have been targeted).

[314] Case C-5/11, Titus Alexander Jochen Donner, ECLI:EU:C:2012:370, ¶¶ 26-30 (June 21, 2012) (sales of copies of protected works amount to copyright infringement where the trader targets the public of the state of destination); Case C-173/11, Football Dataco Ltd. v. Sportradar GmbH, ECLI:EU:C:2012:642, ¶ 27-47 (October 18, 2012) (re-utilization of data within the meaning of Article 7 of Directive 96/9/EC occurs if the act discloses an intention on the part of its performer to target the public in a Member State); Case C-516/13, Dimensione Direct Sales v. Knoll Int'l SpA, ECLI:EU:C:2015:315, ¶¶ 28-35 ((the exclusive distribution right for copyright may be infringed where a trader makes an offer for sale or a targeted advertisement through its website to consumers in another Member State); Case C-597/19, Mircom Int'lContent Mgmt. & Consulting (M.I.C.M.) Ltd. v Telenet BVBA, ECLI:EU:C:2021:492, ¶ 47 (June 17, 2021) (holding that, in the context of peer-to-peer file sharing activities, in order for there to be an 'act of communication,' and consequently, an act of making available, it is sufficient, in the final analysis, that a work is made available to a public in such a way that the persons comprising that public may access it, from wherever and whenever they individually choose, irrespective of whether or not they avail themselves of that opportunity).

[315] Lundstedt, *supra* note 295, at 532-35 (summarizing the case law in United States and EU trademark law and copyright law, with respect to territorial connecting factors for assuming prescriptive jurisdiction, including in particular effects-based criteria).

[316] Rättzén, *supra* note 285, at 372.

[317] WIPO Copyright Treaty art. 8, Dec. 20, 1996, S. Treaty Doc. No. 105-17 (1997); 17 U.S.C. § 106; Infosoc Directive, *supra* note 97, at art.3 (EU); Copyright, Designs and Patents Act 1988, c. 48, § 20 (U.K.).

[318] Rättzén, *supra* note 285, at 372-73.

concurrent jurisdiction. The European Court of Justice has therefore considered that the "mere accessibility" of infringing content alone is not sufficient to amount to trademark[319] or copyright[320] infringement. Instead, the Court found that the substantive test should be whether the infringing activities online are "targeted" to customers within a particular country of protection.[321] This was justified on the basis that, if mere accessibility of a work online from a given country was sufficient to conclude that there was infringement, although obviously targeted at persons outside the territory, then that infringer would "wrongly" be subject to that country's laws.[322] In the United States, courts have so far come to inconsistent conclusions on how to assess when internet use is sufficient to amount to copyright or trademark infringement.[323]

V. LOCALIZATION OF CROSS-BORDER DATA REPRODUCTION FOR MODEL TRAINING

Assuming there has been copying of copyrighted materials in the training dataset or in the encoded parameters of the AI model, and assuming that can be proven, the next question is *where* that infringement would be deemed to have taken

[319] Case C-324/09, L'Oréal SA v. eBay, 2011 E.C.R. I-6073, ¶¶ 61-67.

[320] Case C-173/11, Football Dataco v. Sportradar, ECLI:EU:C:2012:642, ¶¶ 27-47 (October 18, 2012) (accessibility of website is insufficient for performing an act of re-utilization under the sui generis right in Member States). *See also* Case C-5/11, Titus Alexander Jochen Donner, ECLI:EU:C:2012:370, ¶¶ 26-30 (June 21, 2012); Case C-516/13, Dimensione Direct Sales v. Knoll International SpA, ECLI:EU:C:2015:315, ¶¶ 28-35.

[321] *Titus Alexander Jochen Donner*, ECLI:EU:C:2012, ¶¶ 26-30 (sales of copies of protected works amount to an infringement when the trader specifically targets the public of the state of destination); *L'Oréal SA*, 2011 E.C.R., ¶¶ 61-67 (offer for sale of trade-marked goods intended for sale within the EU, or advertising on an online marketplace, must be targeted at consumers in the country of protection); *Dimensione Direct Sales*, ECLI:EU:C:2015, ¶¶ 28-35 (copyright infringement of the exclusive distribution right may be committed when a trader makes an offer for sale or a targeted advertisement through its website to consumers in another Member State where the relevant subject matter is protected by copyright). Relevant factors to determine whether customers are targeted have varied from the language, appearance and content of the website, the nature and size of the business, the characteristics of the goods or services at issue, the content and distribution channels of advertising materials, the number of domestic visitors to the website, and partnerships with shipping companies. *See Titus Alexander Jochen Donner*, ECLI:EU:C:2012, ¶ 29; *See also* Merck KGaA v. Merck Sharp & Dohme Corp [2017] EWCA (Civ) 1834 [170] (Eng.) (summarizing the connecting factors for assessing whether an online advertisement would infringe a trademark).

[322] *Football Dataco*, ECLI:EU:C:2012, ¶ 37.

[323] *See* Perfect 10, Inc. v. Yandex N.V., 962 F. Supp. 2d 1146, 1153-54 (N.D. Cal. 2013) (finding no copyright infringement where it was argued that images hosted on Russian servers could still be downloaded in the United States); L.A. News Serv. v. Conus Communications Co., 969 F. Supp. 579, 583 (C.D. Cal. 1997) (finding copyright infringement where broadcasts were received and viewed within the United States, even if the reception was unintended); Twentieth Century Fox Film Corp. v. iCraveTV, No. CIV.A.00-120, 2000 WL 255989, at *7-9 (W.D. Pa. Feb. 8, 2000) (finding copyright infringement where protected works were made accessible on a streaming website from Canada, even if only Canadian viewers were the intended recipients); Vanity Fair Mills, Inc. v. T. Eaton Co., 234 F.2d 633, 641-43 (2d Cir. 1956) (use of a mark requires a substantial effect on commerce within the United States to amount to trademark infringement).

place. There are two scenarios: first, where the person involved in the training or encoding activity and the server where the data is stored both reside in the same country; and second, where either reside in a different country. Whether intentional or not, the latter scenario is becoming more frequent as developers rely on cloud service providers, whose servers can be located virtually anywhere, for data storage. Similarly, it is possible to relocate the persons, or even the organizations, involved in the training to other countries. For example, a developer may decide to commence its operations and/or place its servers in a country where it is clear that the data collection and processing are not infringing copyright. This becomes a problem for rightholders, however, if that same developer later markets the completed model internationally, including in countries where the same training or processing activity would have been infringing.

The act of reproduction, or copying, is an exclusive right that is territorially limited like all other exclusive rights. If the entire act of copying takes place in a country where there is no infringement, then that is the end of the story for a reproduction infringement case. But the situation becomes more nuanced when the server on which the data is reproduced and stored is located in one country and the person instructing, accessing and using the data is located somewhere else. The reproduction and storage of the data would clearly be an infringing act in the country where the server is located. But it is not as straightforward to say that an infringing act also takes place where the person executes the instruction to download, store and process the data.

This scenario can be compared to management or agent situations, where the management of an organization or a principal instructs its own operations or an agent to commit infringing acts abroad. In *Subafilms*, the Ninth Circuit held that mere authorization within the United States of the distribution of infringing copies abroad did not infringe copyright within the United States.[324] Consistent with earlier decisions, the Ninth Circuit held that the U.S. Copyright Act does not apply to foreign activities.[325] Case law is, however, not conclusive. Years before *Subafilms*, the Second Circuit found in *Update Art* that the U.S. Copyright Act may reach foreign conduct where the type of infringement permits further reproduction

[324] Subafilms, Ltd. v. MGM-Pathe Commc'n Co., 24 F.3d 1088, 1099 (9th Cir. 1994).

[325] *Id.* at 1097.

abroad.[326] Courts in other countries such as Germany[327] and Sweden[328] have also found that mere authorization of foreign infringing acts can amount to an act of infringement domestically. Courts in the U.K.[329] and, inconsistently, some in the United States,[330] have also considered that domestic acts of authorizing foreign infringing acts should be factored in when assessing damages extraterritorially.

The situation resembles that of contributory infringement, where the contributing act takes place domestically, but the primary infringing act takes place abroad. The prevailing view from national courts is that the contributory act in a cross-border situation shall follow the law applicable to the primary infringing act.[331] This means that, if the primary infringing act does not amount to an

[326] Update Art, Inc. v. Modiin Publishing, Ltd., 843 F.2d 67, 73 (2d Cir. 1988). Several district courts have also declined to follow *Subafilms*. *See* Expediters Int'l of Wash., Inc. v. Direct Line Cargo Mgmt. Serv., Inc., 995 F. Supp. 468, 477 (D.N.J. 1998) (holding that "mere authorization of infringing acts abroad constitutes direct infringement and is actionable under United States Copyright Law"); Curb v. MCA Records, Inc., 898 F. Supp. 586, 594 (M.D. Tenn. 1995) ("*Subafilms* relies upon a peculiar interpretation of the scope and nature of the authorization right in 17 U.S.C. § 106. This interpretation . . . appears contrary . . . to well-reasoned precedent, statutory text, and legislative history."); Nat'l Football League v. Primetime 24 Joint Venture, No. 98-3778, 1999 U.S. Dist. LEXIS 3592, at *10 (S.D.N.Y. Mar. 24, 1999) (holding that "where an individual commits an act of infringement in the United States that permits further reproduction outside the United States . . . a court may assert jurisdiction over those foreign acts and a plaintiff may recover damages for the infringing acts that took place extraterritorially"). The ruling in *Subafilms* has also been criticized by scholars. For a summary of the critique, see Nathan R. Wollman, *Maneuvering Through the Landmines of Multi-territorial Copyright Litigation: How To Avoid the Presumption Against Extraterritoriality When Attempting To Recover for the Foreign Exploitation of U.S. Copyrighted Works*, 104 W. VA. L. REV. 343, 374-76 (2002)

[327] *Kreuzbodenventilsäcke*, Bundesgerichtshof [BGH] [Federal Court of Justice] Mar. 29, 1960, Gewerblicher Rechtsschutz und Urheberrecht [GRUR] 423 (Ger.) (finding that an offer in Germany to sell infringing products in another country was considered patent infringement in Germany).

[328] Hovrätt [HovR] [Court of Appeals] 1990 T 1253-89 (Swed.). (finding that a Swedish company's offer for sale of a patented product to a foreign company, delivered from another foreign location, is considered an infringing offer for sale in Sweden). However, in Nytt Juridiskt Arkiv [NJA] [Supreme Court Reports] 2005 p. 180 T 2934-03 (Swed.) [*Formsprutarna*), the Swedish Supreme Court found that the instruction from a Swedish company to manufacture infringing design products in another country did not constitute design infringement.

[329] Experience Hendrix LLC v. Times Newspapers Limited [2010] EWHC (Ch) 1986 [139]-[142] (Eng.) (finding that worldwide damages were suffered as a direct result of the defendant's infringing distribution of sound recordings in the U.K.).

[330] For a summary of the relevant case law on extraterritorial damages awards in the United States, see Timothy R. Holbrook, *Is There a New Extraterritoriality in Intellectual Property?*, 44 COLUM. J.L. & ARTS 457 (2021). In the context of patent law, see also Thomas F. Cotter, *Extraterritorial Damages in Patent Law*, 39 CARDOZO ARTS & ENT. L.J. 1 (2021).

[331] Subafilms, Ltd. v. MGM-Pathe Commc'n Co., 24 F.3d 1088, 1092 (9th Cir. 1994) (holding that "there could be no liability for contributory infringement unless the authorized or otherwise encouraged activity itself could amount to infringement"); Abkco Music & Records Inc. v. Music Collection International Limited [1995] RPC 657 at 660-661 (Eng.) (holding that the act of authorization does not have to occur in the U.K., provided that the primary act of infringement so authorized does); *Folgerecht bei Auslandsbezug*, Bundesgerichtshof [BGH] [Federal Court of Justice] June 16, 1994, Gewerblicher Rechtsschutz und Urheberrecht, Internationaler Teil [GRUR] 798 (Ger.) (holding that the act of authorization in Germany to conduct an auction of infringing products in another country did not amount to an infringement in Germany).

infringement abroad, then neither can the contributory act be infringing at home. In line with that premise, Mr. Justice Arnold recently held in *British Broadcasting* that "[c]onsistent[] with the territorial nature of U.K. copyright, any act constituting a primary infringement of copyright must take place within the U.K."[332]

The issue in the context of model training would stem from an act of reproduction that is wholly completed abroad. There is no doubt that copies are physically reproduced on servers where they are downloaded and stored. If the server is located abroad, then there is simply no additional territorial connection in the act of authorizing that reproduction. This arguably makes it different from other types of exclusive rights, such as offering for sale or communicating to the public, where there are clearly two elements and two actors: the act itself performed by one actor and the recipient of that other act. Both of these elements can have their own territorial connections to different countries and infringe in both countries or in either or neither of the two. While it is possible in theory to consider that instructing the reproduction of infringing datasets on servers located abroad is a contributory infringing act, that does not leave rightholders in a better situation. The primary infringing act is still the actual copying occurring on the server, and if those servers reside in a country where there is no infringement, then logically, the contributory act cannot infringe either.

The issue of where training activities had taken place recently came into the spotlight before the English High Court in *Getty Images v. Stability AI*.[333] In that case, Getty Images argued that Stability AI had infringed its rights by using Getty's images as data inputs for the purposes of training and developing its image generation model, Stable Diffusion. Stability AI, which is a U.K. company, alleged that it had committed no infringing acts in the U.K.[334] Its defense relied on the fact that its training activities took place outside the U.K. and, therefore, that there could be no infringement in the U.K.[335] In particular, it pleaded that "[n]one of the individuals who were involved in developing and training Stable Diffusion . . . resided or worked in the U.K. at any material time during its development and training."[336] It then went on to plead that: "[t]he development work associated with designing and coding the software framework for Stable Diffusion and developing a code base for training it was carried out . . . outside the U.K. The training of each iteration of Stable Diffusion was performed . . . outside the U.K. No visual assets or associated captions were downloaded or stored (whether on servers or local devices) in the U.K. during this process."[337] Mrs. Justice Smith of the High Court

[332] Broadcasting Corporation & Anor v. Mechanical-Copyright Protection Society Ltd & Ors, [2018] EWHC (Ch) 2931 [26] (Eng.).

[333] Getty Images (US) Inc v. Stability AI Ltd [2023] EWHC (Ch) 3090.

[334] *Id.* [55]-[56].

[335] *Id.* [11].

[336] *See* Cerys Wyn Davies & Gill Dennis, *Getty Images v Stability AI: the implications for U.K. copyright law and licensing, Pinsent Mason* (Apr. 29, 2024), https://www.pinsentmasons.com/out-law/analysis/getty-images-v-stability-ai-implications-copyright-law-licensing [https://perma.cc/Y4LK-5XP7] (citing Stability AI's defense arguments from the pleadings).

[337] *Id.*

briefly commented on the issue when considering Stability AI's failed strike-out application in December 2023, stating that "if this were the trial of this action, the evidence to which I have referred above would (on its face) provide strong support for a finding that, on the balance of probabilities, no development or training of Stable Diffusion has taken place in the United Kingdom."[338] Because of that, the Court permitted the defense to proceed to be argued in full at trial.[339]

What makes the enforcement situation so difficult for rightholders is that servers, of course, can be situated almost anywhere with little effort. If courts are to find infringement in their forum in these circumstances, they would have to adopt far-reaching statutory readings of what counts as reproduction and focus on maintaining the efficacy of copyright laws to avoid loopholes. Historically, and as discussed above, some courts have shown a willingness to extend the reach of intellectual property laws where infringing conduct can be arbitrarily located somewhere else,[340] particularly when the infringing party can still benefit from that foreign conduct in the forum.[341] This is no doubt the case for AI model training. A model can be trained on servers located in one country, and then, once complete, be marketed and sold in another country without restrictions.

VI. A TRANSNATIONAL DATA LOOPHOLE FOR MODEL TRAINING

The discussion above focused on the situation where AI developers decide to relocate their servers to a first country and maintain their business operations and staff in a second country. Courts in different jurisdictions are expected to be more or less sympathetic to finding that an act of reproduction takes place in the second country in these circumstances. However, rightholders will face an even greater uphill battle where AI developers have decided to relocate *both* their servers and their staff and business operations to more AI-friendly countries with exceptions and limitations for model training. The review of copyright exceptions and

[338] Getty Images (US) Inc v. Stability AI Ltd [2023] EWHC (Ch) 3090 [59].

[339] *Id.* [60]. The case is presently scheduled to be heard at trial in the summer of 2025. *How will the new U.K. government resolve the conflict over AI development and IP rights?*, OSBORNE CLARKE (July 7, 2024), https://www.osborneclarke.com/insights/how-will-new-uk-government-resolve-conflict-over-ai-development-and-ip-rights [https://perma.cc/4UBP-JQWQ].

[340] *See* Case C-324/09, L'Oréal SA v. eBay, 2011 E.C.R. I-6073 ¶¶ 61-62 (finding that EU trademark rules should apply for goods located outside the EU if their online sale is targeted at EU consumers); Case C-173/11, Football Dataco v. Sportradar, ECLI:EU:C:2012:642, ¶¶ 44-47 (October 18, 2012) (defining the act of sending data from a web server in Member State A to a computer in Member State B, at the request of the recipient, as 're-utilisation' of the data, particularly when there is intent to target the public in Member State B). Similarly, in the patent context, the England and Wales High Court held in *Illumina, Inc v. Premaitha Health Plc* that there would be direct infringement in the U.K. for a method patent claim involving blood diagnostics, where the method was carried out abroad where there was no patent and where the results from the method were subsequently provided to customers in the U.K. [2017] EWHC (Pat) 2930 [507]-[508] (finding that the "substance" of the method was still performed inside the U.K. as blood samples were retrieved from U.K. customers and that "any other result would make it far too easy to avoid infringement of patents of this nature, given the ease of digital transmission and the ability to off-shore computer processing.").

[341] *See* cases cited *supra* note 312 and accompanying text.

limitations in different countries, discussed previously in this Article,[342] shows that this is not a mere hypothetical possibility. There clearly are considerable differences between different countries in assessing whether the reproduction of training data, either when creating the original dataset or when encoding the data into model parameters, amounts to copyright infringement. If developers choose to relocate both their training activities and their business operations to a country where this is permitted, then it is unlikely that there will be *any* infringement *anywhere*. The act of reproduction occurs entirely abroad in this scenario. There will be no connecting factors, let alone any "sufficient" connecting factors, relating to the data reproduction that can be localized to any other countries.

However, developers may still later decide to make their completed model available to customers in other countries where the training activities would have been infringing *if* they had occurred there. It would be unsatisfactory to rightholders if they are denied legal recourse in their respective countries, which have taken the position that copyright protection for original content should prevail merely because the data reproduction activities in training the model took place somewhere else. Copyright protection becomes worthless if those developers could subsequently sell and market their completed, pre-trained AI models or if custom developers could sell and market their fine-tuned AI models or AI systems based on those pre-trained models in other countries where the training process would be infringing. Furthermore, what many rightholders are concerned about is not so much the reproduction of their works during the training phase. What is arguably most threatening to rightholders are the generated outputs produced once the AI model is complete. It is the completed model and its output, not the process behind it and its input, that could reduce demand for original copyrighted works.[343] While derivative works may provide some protection to rightholders in certain contexts, this will generally be limited depending on the circumstances and provides no blanket solution.[344]

This situation marginally leaves rightholders with one remaining option as copyright law currently stands, which is to focus on the infringing acts that can be attributed to the end use of the AI model or AI systems based on the model. This tactic can be successful where the AI generates works that are identical or substantially similar to the original works, for example because the model has "memorized" parts of the training data. Like any other works that meet the relevant infringement standard, these AI-generated works will directly infringe the copyright of authors whose works have been used as training data. However, it will be the exception rather than the norm that AI models "memorize" the training data and are capable of replicating it identically or near-identically when prompted by

[342] *See supra* Sections III.B, III.C.3.

[343] *See supra* Section III.B.1 (discussing, in relation to fair use, how AI-generated works can reduce the demand of original authors' works). *See also infra* Section VII.C and note 440 (discussing how these arguments have been framed by plaintiffs in recent copyright complaints).

[344] The issue of derivative works is discussed separately below, *see infra* Section VIII.A.

the end user.[345] The vast majority of AI-generated works will not be directly infringing the training data. They may, bit by bit, draw upon statistical representations and patterns from a very large number of different works, or they may only be moderately similar or indicative of any original works or the style in such works. This is unlikely to be copyright infringement.

If rightholders cannot enforce their rights in the course of model training because developers have relocated their training activities to AI-friendly jurisdictions and if rightholders cannot enforce their rights in relation to the AI-generated output, they are left without options. This ultimately creates a transnational loophole in the international copyright system as it stands, leaving rightholders with little to no protection from their works being used against their will to develop AI models that reduce the demand for their original creations. There is currently only one thing that could stop this from happening, and that is if copyright laws are applied extraterritorially.

VII. Closing the Transnational Data Loophole

A. An Extraterritorial Application of Copyright Laws in the Context of AI

1. The EU AI Act and its Extraterritorial Extension of EU Copyright Law

Regulators have unsurprisingly picked up on this problem. In the EU, the Copyright Directive does not expressly state whether developers conducting their data mining operations in third countries must also follow the same opt-out procedure set out in Article 4(3), as would clearly be the case if the training took place within the EU. In an afterthought, the EU regulators have sought to extend the extraterritorial reach of the opt-out procedure. Recital 106 of the EU AI Act, which was recently enacted in the EU, now states that "[a]ny provider placing a general-purpose AI model on the Union market should comply with [Article 4(3) of the Copyright Directive], regardless of the jurisdiction in which the copyright-relevant acts underpinning the training of those general-purpose AI models take place." This was necessary, it was explained, "to ensure a level playing field among providers of general-purpose AI models where no provider should be able to gain a competitive advantage in the Union market by applying lower copyright standards than those provided in the Union."[346] In other words, what the regulators were concerned about was the situation where providers of general-purpose AI outside the EU are put in a more favorable competitive position than those established in

[345] *See supra* Section III.C.1, referring to Somepalli et al., *supra* note 253, at 2 (finding in a survey, testing Stable Diffusion v1.4, that about 1,2% of 100,000 randomly sampled user-generated captions were potentially sufficiently similar to the original training data, indicating that they may be duplicates). For further studies conducted on the issue of memorization in AI models, which suggest that this is a rare technical occurrence except when fine-tuned datasets are employed, see *infra* notes 481-483. The issue of targeting the end use of generative AI models for copyright infringement, thereby avoiding the issue of copyright territoriality, is discussed in more detail in *infra* Section VIII.

[346] EU AI Act, *supra* note 11, at Recital 106.

the EU if they do not have respect for rightholders' opt-outs. This would mean that such foreign providers could obtain larger datasets without any regard for copyrighted materials and still compete on the same market in the EU. While that is a sound reason, it is a very different justification than that of concern for rightholders' possibility to enforce their rights when the infringing conduct has occurred outside the EU. It should be noted that the recitals in EU acts are not legally binding and have no independent legal value, but rather, interpretative value in that they are capable of explaining provisions and clarifying the intention of the authors.[347] It is highly unusual that recitals address how other acts should be interpreted, which, of course, were drafted by different authors and under different circumstances. It is also highly unusual that recitals address important issues such as extraterritorial scope.

Article 53.1(c) of the EU AI Act may provide a similar reading to Recital 106. In particular, it sets out that providers of general-purpose AI models shall "put in place a policy to comply with Union copyright law, and in particular to identify and comply with, including through state-of-the-art technologies, a reservation of rights expressed pursuant to Article 4(3) of Directive (EU) 2019/790." There is no territorial limitation for that obligation to comply with rightholders' opt-out procedure only in cases where the training has been conducted in the EU. Article 2 of the EU AI Act generally broadens the scope of the Act to providers of general-purpose AI models when placing their models on the European market, irrespective of whether those providers are established or located within the EU or in a third country.[348] Therefore, providers of general-purpose AI models who have developed the model in a third country outside the EU will still need to comply with the entirety of the Act, including Article 53.1(c). Because of that, there is support for saying that the extraterritorial reach of complying with the opt-out procedure set out in Article 4(3) of the Copyright Directive is not only limited to the recitals of the Act.

Both Recital 106 and Article 53(1)(c) of the EU AI Act mean that foreign developers who conduct their data mining activities outside the EU have to comply, extraterritorially and retroactively, with any opt-outs from rightholders if and when they later decide to place their general-purpose AI models on the European market. This closes the transnational data loophole in the sense that providers of general-purpose AI models cannot escape the requirement of respecting opt-outs from rightholders no matter where the training activities took place. Yet both Recital 106 and Article 53(1)(c) of the EU AI Act leave many questions open, and, upon closer examination, it is unclear if it really closes any loophole at all.

First, the rules are limited to "general-purpose AI models," which refer to AI models trained with a large amount of data using self-supervision at scale, that display significant generality, that are capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market, and

[347] Case C-755/21 P, Marián Kočner v. Europol, ECLI:EU:C:2024:202, ¶ 59 (March 5, 2024).

[348] EU AI Act, *supra* note 11, at art. 2(1)(a).

that can be integrated into a variety of downstream systems or applications.[349] That definition is significantly narrower than Article 4(3) of the Copyright Directive, which generally applies to "text and data mining" operations including any automated analytical technique aimed at analyzing text and data in digital form in order to generate information.[350] This raises the question whether Article 4(3) should be applied extraterritorially in all circumstances, or merely for general-purpose AI models to which the EU AI Act applies. The express limitation in the EU AI Act suggests that it is only the latter.

Second, because Recital 106 and Article 53(1)(c) of the EU AI Act are textually limited to general-purpose AI models, other AI systems will not need to comply with EU copyright law if the training activities are conducted outside the EU. The EU AI act makes an important distinction between general-purpose AI models, AI systems, and general-purpose AI systems. AI systems are broadly defined as machine-based systems that are designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infer, from the input they receive, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.[351] General-purpose AI systems are also defined as AI systems, but more narrowly a general-purpose AI model, which has the capability to serve a variety of purposes, both for direct use as well as for integration in other AI systems.[352] Leaving those definitions aside, because Recital 106 and Article 53.1(c) of the EU AI Act are limited specifically to providers of general-purpose AI models only, providers of AI systems or general-purpose AI systems do not have to extraterritorially comply with EU copyright law. However, if those providers are based within the EU and carry out the training activities there, they will need to comply with Article 4(3) of the Copyright Directive.

This discrepancy turns into a very awkward legal situation and precisely what the EU regulator sought to avoid when Recital 106 stated that it was "to ensure a level playing field among providers of general-purpose AI models." Why should there be a level playing field from a copyright perspective among providers of general-purpose AI models but not of AI systems generally? Examples of model types that are clearly captured by the former definition include LLMs,[353] large multimodal models ("LMMs") and other foundation models. But AI models which are not general-purpose AI models can, of course, also be trained on significant

[349] *Id.* at art. 3(63).

[350] Copyright Directive, *supra* note 117, at art. 2(2).

[351] EU AI Act, *supra* note 11, at art. 3(1).

[352] *Id.* at art. 3(66). *See also id* at Recital 100, explaining that "[w]hen a general-purpose AI model is integrated into or forms part of an AI system, this system should be considered to be general-purpose AI system when, due to this integration, this system has the capability to serve a variety of purposes. A general-purpose AI system can be used directly, or it may be integrated into other AI systems."

[353] *Id.* at Recital 99 (stating that "[l]arge generative AI models are a typical example for a general-purpose AI model, given that they allow for flexible generation of content, such as in the form of text, audio, images or video, that can readily accommodate a wide range of distinctive tasks.").

amounts of data using self-supervision. Large datasets are routinely required for machine learning and AI applications in general. It remains unclear why copyright provisions were only included in the EU AI Act with respect to general-purpose AI models only.

To make matters worse, the fact that providers of general-purpose AI systems, instead of models, are *not* obliged to extraterritorially follow Article 4(3) of the Copyright Directive could create another "loophole" in itself. General-purpose AI systems are AI systems based on general-purpose AI models, which would include fine-tuned versions of general-purpose AI models or other custom-built versions of such models. Because only providers of general-purpose AI models, but not systems, have to comply with this obligation, it is theoretically possible to structure business operations such that the provider never themselves places the general-purpose AI model on the European market. Instead, they could provide the general-purpose AI model to only custom developers, located outside Europe, who in turn provide their general-purpose AI systems to the EU. Such providers of general-purpose AI systems would not need to comply with the same extraterritoriality provision, at least on the face of the text. If this understanding is correct, then neither Recital 106 nor Article 53.1(c) really closes a transnational data loophole at all. In that case, it would be possible to avoid the obligation to respect rightholders' opt-outs by offshoring the relevant training activities and segregating business operations such that the general-purpose AI model itself is never placed directly on the EU market. This arguably goes against the spirit of Recital 106 and Article 53.1(c) to "ensure a level playing field" for the use of general-purpose AI models within the EU.[354] But clearly, that objective has already failed by excluding AI systems from the same obligations.

Perhaps the only way to close this regulatory loophole without doing a logical somersault would be to broadly interpret when someone is deemed to "provide" a general-purpose AI model on the EU market. "Providers" in this regard are defined as: any natural or legal person, public authority, agency, or other body that develops a general-purpose AI model or AI system or that has a general-purpose AI model or AI system developed and "places it on the market" or "puts it into service" under their own name or trademark, whether for payment or free of charge.[355] A general-purpose AI model or AI system is "placed on the market" where it is first "made available on the market" in the EU,[356] and "put into service" where it is supplied for first use directly to the deployer or for own use in the EU for its intended purpose.[357] The former is broader than the latter. The "making available on the market" means the supply of a general-purpose AI model or AI system for

[354] *Id.* at Recital 106, art. 53(1)(c). It would be difficult to rectify this logical gap through a purposive interpretation of Recital 106 and Article 53(1)(c). Both of these rules clearly only apply to providers of general-purpose AI models in the wording, and it would be a very significant leap, inconsistent with ensuring legal certainty, to say that also different actors and different AI systems should follow the same rules as well.

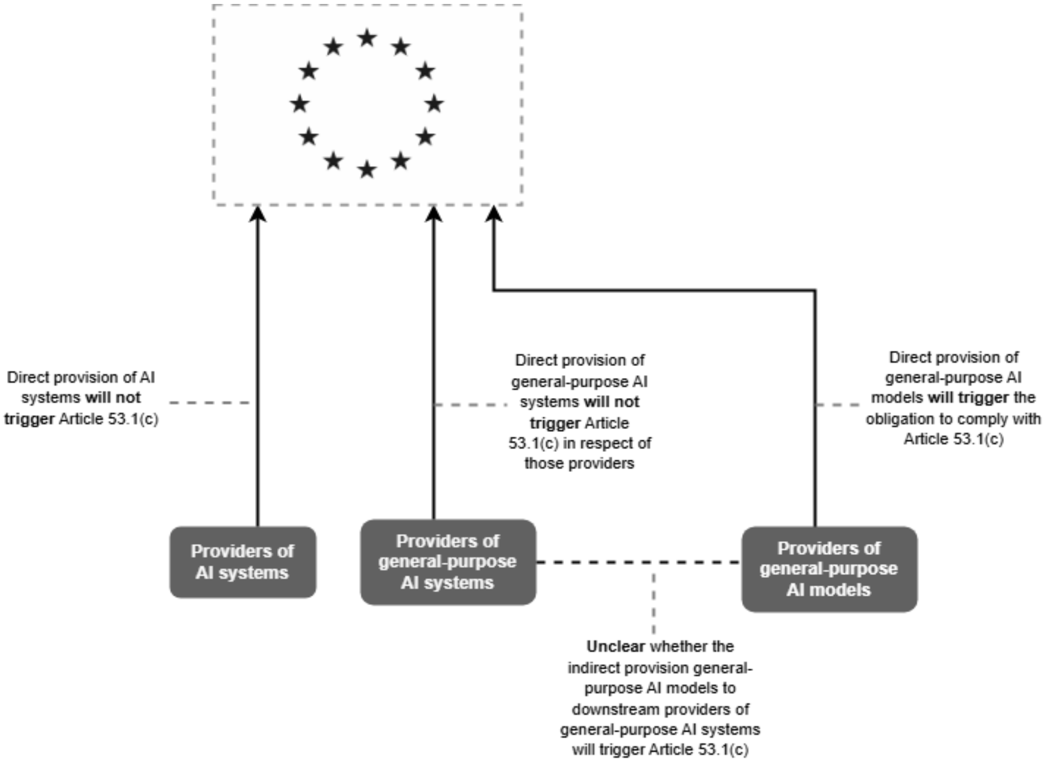
[355] *Id.* at art. 3(3).

[356] EU AI Act, *supra* note 11, at art. 3(9).

[357] *Id.* at art. 3(11).

distribution or use in the EU in the course of a commercial activity and does not require that there is any actual recipient using the model or system.[358] With this in mind, it might be argued that providers of general-purpose AI models also make their models available on the EU market through downstream suppliers, who in turn provide general-purpose AI systems based on the model in the EU even if the model is never directly placed on the EU market. Otherwise, the obligations in the EU AI Act for providers of general-purpose AI models, including Recital 106 and Article 53.1(c), could easily be circumvented.

Figure 7: Overview of when Article 53.1(c) of the EU AI Act applies extraterritorially


Direct provision of AI systems **will not trigger** Article 53.1(c)
Direct provision of general-purpose AI systems **will not trigger** Article 53.1(c) in respect of those providers
Direct provision of general-purpose AI models **will trigger** the obligation to comply with Article 53.1(c)
Providers of AI systems
Providers of general-purpose AI systems
Providers of general-purpose AI models
Unclear whether the indirect provision general-purpose AI models to downstream providers of general-purpose AI systems will trigger Article 53.1(c)

Third, another ambiguity relating to Recital 106 and Article 53.1(c) is whether they actually extend extraterritorially the reach of Article 4(3) of the Copyright Directive. Strictly speaking, the obligation to put in place a copyright policy that respects rightholders' opt-outs forms part of the EU AI Act. This also means that, strictly speaking, failing to put in place a policy that complies with EU copyright law would not amount to copyright infringement and therefore would not lead to any remedies in copyright law but only sanctions for non-compliance of the EU AI Act.[359] This is a subtle, but very important, distinction. Although the penalties of not complying with the EU AI Act are significant and have a deterrent effect on

[358] *Id.* at art. 3(10).

[359] *Id.* at art. 99.

their own,[360] any administrative fines would come to benefit public authorities instead of rightholders. This reinforces the point already made above, that the extraterritorial application of EU copyright law in the context of the EU AI Act is *not* for the benefit of rightholders but to even out competition among AI developers in the EU. However, this is at the same time illogical because Recital 107 explains at the same time that the obligation to draw up a summary of the content used for training general-purpose AI models is meant to "facilitate parties with legitimate interests, including copyright holders, to exercise and enforce their rights under Union law." It is wholly unclear why rightholders should bother to look through summaries of training datasets to see if their works have been included even if they have opted-out of this if they have no enforcement mechanisms against such infringements. Because of that logical gap, the other way of reading Recital 106 and Article 53.1(c) is that, as a matter of copyright law, they extraterritorially extend the ambit of the opt-out procedure in the Copyright Directive. Developers who have trained and developed their models in foreign jurisdictions and then rolled out their models for use within the EU would be required to produce a summary to EU authorities about the content used for model training.[361] If developers have not complied with the opt-out procedure set out in Article 4(3) of the Copyright Directive for their foreign training activities, then they might be liable for *both* copyright infringement and penalties under the EU AI Act, which together could become very substantial. Ultimately, it will no doubt fall on the task of the European Court of Justice to resolve this serious ambiguity.

Fourth, if read narrowly, Recital 106 and Article 53(1)(c) merely establish an obligation to "put in place a policy" to comply with rightholders' opt-outs when using copyrighted materials for the purpose of text and data mining. This involves drawing up and making publicly available a sufficiently detailed summary of the data used for training the general-purpose AI model.[362] Although it is not clear, this presumably also requires describing in the copyright policy how rightholders' opt-outs are respected, such that works from rightholders who have reserved their rights for text and data mining are excluded from the training dataset. However, this is still just an obligation to "put in place a policy" to do so. What happens if the provider of a general-purpose AI model puts in place a sufficiently detailed policy, detailing how it aims to comply with EU copyright law, including Article 4(3) of the Copyright Directive but failing to respect rightholders' opt-outs? On the face of

[360] Failure to comply with the EU AI Act can result in administrative fines of up to €15 million, or up to 3% of total worldwide annual turnover for the preceding financial year, whichever is higher. *See id.* at art. 101(1)(a).

[361] *See* EU AI Act, *supra* note 11, at art. 53(1)(d). *See also* Copyright Directive, *supra* note 117, at Recital 104 ("[G]iven that the release of general-purpose AI models under free and open source license does not necessarily reveal substantial information on the data set used for the training or fine-tuning of the model and on how compliance of copyright law was thereby ensured, the exception provided for general-purpose AI models from compliance with the transparency-related requirements should not concern the obligation to produce a summary about the content used for model training and the obligation to put in place a policy to comply with Union copyright law, in particular to identify and comply with the reservation of rights pursuant to Article 4(3) of Directive (EU) 2019/790 of the European Parliament and of the Council.").

[362] EU AI Act, *supra* note 11, at art. 53(1)(c), Recital 107.

the text of Recital 106 and Article 53(1)(c), the provider has done everything it should have done and nothing more. However, if this is the correct interpretation, it would mean that the obligation to put in place a policy would have little to no force if that policy did not, as a matter of substance, also actually comply with EU copyright law. It is well-settled that European acts are purposively interpreted to give effect to the aim or spirit of the legislation,[363] which favors a broader interpretation.

Fifth, regardless of which interpretation is adopted, a practical hurdle for developers would be timelines for retroactive copyright compliance. Those who have not yet placed their general-purpose AI models on the EU market have twelve months to comply from the date of entry into force of Act,[364] while those who already have placed their general-purpose AI models on the EU market before that time have thirty-six months to comply.[365] This is easier said than done if models will need to be retrained on more limited datasets after having sifted out from the dataset any works where rightholders have opted-out from text and data mining.

Last but not least, it is questionable whether the underlying assumption in Recital 106, to ensure a level playing field among providers of general-purpose AI models, holds enough merit to warrant an extraterritorial application for all cases. A developer could train their model in another country with *different* rather than *lower* copyright standards than those provided in the Copyright Directive. This would be the case in the United States, where fair use is assessed on a case-by-case basis. A developer could also be training their model in another country with *higher* copyright standards. But in both of these situations, Recital 106 would still require that foreign developers adhere to the distinct opt-out procedure set out in the Copyright Directive. It is difficult to justify such a blanket approach to extraterritoriality and makes compliance unnecessarily burdensome and complex for developers. One way to make sense of all of this in a more balanced way is to interpret Recital 106 narrowly, such that the extraterritoriality obligation *only* applies when the training activities take place in a country with demonstratively *lower* copyright standards than those in the EU. However, this leaves us with the difficult, and even political, question of trying to answer what are “lower,” “normal,” and “higher” standards of copyright protection.

2. International Trends Towards Copyright Extraterritoriality for AI

It is not uncommon to find similar provisions in other EU acts having extraterritorial reach. For example, similar provisions can be found more generally

[363] Timo Rademacher, *Reading Up or Down EU Legislation: A Plea for a Principled Approach to an Extraordinary Judicial Power*, 23 EPL 319 (2017). *See also Brent London Borough Council and others v. Risk Management Partners Ltd* [2011] UKSC 7, at ¶ 25 (advocating for a purposive interpretation of U.K. regulations based on EU law).

[364] EU AI Act, *supra* note 11, at art. 113.

[365] *Id.* at art. 111(3).

in the EU AI Act,[366] and as discussed above, and the GDPR,[367] requiring that foreign entities that place their products or services on the European market meet the necessary requirements. Because the European market comprises hundreds of millions of consumers with considerable purchasing power, access to the market is essential for companies with international operations. This has led to the so-called Brussels effect, which influences global regulatory standards by requiring all companies to adopt more stringent EU standards to access the large European market.[368] Foreign companies that do some business in the EU will then need to comply with the EU rules. The EU has taken a pioneering role as the first in the world to implement comprehensive regulations for AI. These rules are likely to have a significant impact on foreign developers interested in accessing the European market as well as foreign regulators by serving as a model for AI regulations.

It remains uncertain whether this following trend will extend to copyright aspects. Historically, copyright law has varied considerably across different countries, reflecting diverse economic, social, and cultural factors unique to each region. The comparative survey above showcased that countries so far are taking markedly different approaches when it comes to balancing AI with copyright law,[369] and many regulators have adopted a "wait-and-see" approach, including the U.K., Australia, and India.[370] There is currently no sign that other countries will be content with merely following an opt-out procedure for rightholders to solve the copyright infringement dilemma. Indeed, some countries, such as Israel and Singapore[371] have already taken an AI-friendly approach, which means that most text and data mining activities will be permitted and not amount to copyright infringement. In other countries, such as the United States[372] and Japan,[373] whether data reproduction for model training and encoding amounts to copyright infringement will be a case-by-case analysis. To be fair to the EU in this regard, the opt-out procedure set out in the Copyright Directive was negotiated and agreed upon between Member States under different circumstances, *in casu*, before the AI breakthroughs in recent years. It was also initially meant to tackle the issue of transaction costs and legal uncertainty for research organizations[374] which were

[366] *Id.* at art. 2(1).

[367] GDPR, *supra* note 13, at art. 3(2).

[368] *See* Anu Bradford, THE BRUSSELS EFFECT: HOW THE EUROPEAN UNION RULES THE WORLD 2020.

[369] *See supra* Section III.B and Section III.C.3.

[370] As discussed in *supra* Section III.B.4, most countries do not have copyright exceptions and limitations for text and data mining. Meanwhile, these questions are being actively debated at an international level at WIPO.

[371] *Id.*

[372] *See supra* Section III.B.1.

[373] *See supra* Section III.B.4.

[374] *See supra* Section III.B.2 (discussing how the text and data mining exception in the Copyright Directive, *supra* note 117, was initially focused on research organizations).

conducting text and data mining activities on a more limited basis than what is the case for AI developers today.

Ultimately, it remains to be seen whether the opt-out compromise in the Copyright Directive will be effective in balancing both the interests of rightholders and AI providers in the new era of generative AI. Although it is "nice in thought," if too many rightholders decide to opt-out of text and data mining, it could become impractical for developers who might run into data scarcity issues unless licenses are procured for accessing more content which would further drive development costs. Extended collective licensing arrangements, such as the one recently proposed in Denmark,[375] could effectively fill these gaps without the need for procuring individual licenses. It equally remains to be seen whether more Member States within the EU will introduce extended collective licensing arrangements for AI training activities.[376] One issue with collective licensing is that it is territorial in nature.[377] Either all Member States have to introduce similar arrangements in national legislation for the AI context, or new EU legislation is needed to provide for multi-territorial collective licensing specifically in the context of text and data mining. The latter would then become similar to what happened for online use of musical works.[378] The other issue with collective licensing is that it is a legal innovation that is more common in the EU, particularly Scandinavia. It is not widespread globally and lacks international harmonization.

Another example of regulations with extraterritorial reach which has recently come to the spotlight is Brazil, where new AI regulations are currently pending review.[379] Similar to the position taken in the EU Copyright Directive, the latest Draft Brazilian AI Regulations exclude data and text mining related processes in AI systems from constituting copyright infringement under certain conditions.[380] More specifically, for data and text mining to be excluded from copyright infringement, access to the original work must have been lawful and not have as its main objective the reproduction, exhibition or dissemination of the original work itself.[381] Use of protected content must also be necessary for the objective to be achieved, and must not unreasonably prejudice the economic interests of

[375] *See supra* Section III.B.2 (under the heading "Minimum Harmonization and Extended Collective Licensing").

[376] *See* Copyright Directive, *supra* note 117, at art. 12 (Member States now have an express mandate to do so under Article 12 of the Copyright Directive).

[377] *See* Copyright Directive, *supra* note 117, at Recital 46. (stating that, "[s]uch [collective licensing] mechanisms should only have effect in the territory of the Member State concerned, unless otherwise provided for in Union law."). The territorial nature of extended collective licensing comes from the fact that the copyright laws which grant such rights are territorial themselves.

[378] *See* Council Directive 2014/26, 2014 O.J. (L 84) 72 (EU) on collective management of copyright and related rights and multi-territorial licensing of rights in musical works for online use in the internal market.

[379] Projeto de lei No. 2338/2023 [Bill No. 2338/2023], Senado Federal, Sessão Legislativa de 2023 (2023) (Braz.), https://legis.senado.leg.br/sdleg-getter/documento?dm=9852013&ts=1733427140776&rendition_principal=S&disposition=inline [hereinafter Draft Brazilian AI Regulations].

[380] *Id.* at art. 63.

[381] *Id.* at art. 63(I)(-II).

rightholders and not compete with the normal exploitation of the works.[382] However, and importantly, the copyright exception only applies in the current draft to research, journalism, museum, archive, library and educational organizations and institutions and is limited to text and data mining processes that do not have commercial purposes.[383] The regulations do not define "research organizations and institutions," and therefore it is not entirely clear whether private entities could also enjoy the exception. However, because text and data mining processes that have commercial purposes are excluded from the exception accord, it would appear that the answer is negative. This is confirmed by Article 63, second paragraph, of the regulations, which exclude the application of the exception to "institutions linked, affiliated or controlled by a for-profit entity that provides or operates AI systems." The default position under Article 64 is then that rightholders may prohibit the use of their protected contents in the development of AI and that such rightholders may request remuneration from AI agents under Article 65.[384]

The pending AI regulations in Brazil, if unchanged and enacted, would apply to "developers" developing an AI system with the intention of "placing it" on the Brazilian market or "applying it in a service provided" to Brazilian users, for payment or free of charge, as well as to "applicators" who "employ or use" an AI system in Brazil.[385] "AI agents" are defined in Article 4(VIII) as encompassing both developers, distributors, and applicators that operate in the value chain and internal governance of AI systems. The Brazilians draft rules are similar in this regard to the extraterritorial scope of Articles 2(a) and 2(c) of the EU AI Act. In practice, this could mean that foreign AI developers, if constituting research organizations and institutions, would have to comply with the Brazilian copyright exceptions and limitations on data and text mining if they market or sell their model to customers or users in Brazil even *if* their training and development activities take place in another country and even *if* they are not infringing there. What is unique about the Brazilian proposal compared with Recital 106 of the EU AI Act, is that it contains explicit provisions about copyright infringement and, therefore, clearly extends their reach to exclusively foreign training activities. The situation is less clear with respect to foreign AI developers who do not constitute research organizations and institutions, or who conduct text and data mining for commercial purposes and hence do not satisfy the conditions in Article 63. Then, the copyright exceptions and limitations will not apply in the current draft of the Brazilian draft regulations, and there is instead an obligation in Article 65 for "AI agents" to remunerate

[382] *Id.* at art. 63(IV). This is similar to, but not the same as, the three-step test in Article 9(2) of the Berne Convention, *supra* note 92, and Article 13 of the TRIPS Agreement, *supra* note 130.

[383] *Id.* at art. 63(II).

[384] Article 65 of the Draft Brazilian AI Regulations provides that "[t]he AI agent that uses content protected by copyright and related rights in processes of mining, training or development of AI systems must remunerate the respective holders of such content by virtue of such use." However, and interestingly, this right to remuneration only applies to holders of national or foreign copyright and related rights domiciled in Brazil, or to persons domiciled in another country that ensures reciprocity in copyright protection, in terms equivalent to the Draft Brazilian AI Regulations. *See* art. 65(V).

[385] *Id.* at art. 4(V) and (VII).

rightholders. It is not clear whether this obligation also extends extraterritorially. If answered negatively, then foreign training activities, if having no connecting factors to Brazil that could make them directly infringing acts of reproduction, will likely *not* infringe on Brazilian copyright.[386] If that assumption is correct, then the obligation for AI agents, who are not research organizations and institutions, to remunerate rightholders under Article 65 of the draft regulations would likely also *not* apply to foreign training activities. This would altogether make the Brazilian draft proposal largely ineffective in closing the transnational data loophole that currently exists.

The country that may be one of the least likely to adopt an extraterritoriality approach to copyright law is the United States. Historically, courts in the United States have been cautious about extending the reach of intellectual property law to foreign conduct.[387] This is grounded in the presumption against extraterritoriality, a legal concept which uniquely exists in the United States[388] and which functions as a canon of statutory interpretation. In the absence of a contrary, clear, and affirmative statement from Congress, federal laws are presumed to operate only within the territorial jurisdiction of the United States.[389] The presumption has been justified on the basis that "Congress ordinarily legislates with respect to domestic, not foreign matters"[390] and to "avoid unreasonable interference with the sovereign authority of other nations."[391] The chances that such courts would, without any legal basis, extraterritorially extend the U.S. Copyright Act to entirely foreign acts of reproduction are therefore slim. It is instead Congress that must act if anything should happen with any certainty.

3. How Copyright Extraterritoriality Could Undermine the AI Industry

Both the far-reaching statement in the recitals of the EU AI Act and Draft Brazilian AI Regulations portend something problematic. As regulators around the world push forward new regulations impacting AI, they so far appear to be doing so extraterritorially out of fear that developers, as OpenAI itself has admitted,[392]

[386] This follows as a result of the transnational data loophole that exists due to the nature of copyright territoriality. *See supra* Section VI. In contrast, if the Brazilian copyright rules would be interpreted in such a way to apply extraterritorially to foreign training activities, then this would arguably conflict with foreign sovereignty and Article 5(2) of the Berne Convention, *supra* note 92. *See infra* Section VII(A)(4).

[387] The *Subafilms* case from the Ninth Circuit is a good example of that judicial restraint. *See* Subafilms, Ltd. v. MGM-Pathe Commc'n. Co., 24 F.3d 1088, 1099 (9th Cir. 1994). *See also* discussion *supra* Section V.

[388] Subafilms, 24 F.3d at 1095 (citing the presumption against extraterritoriality).

[389] EEOC v. Arabian American Oil Co., 499 U.S. 244, 248 (1991); Morrison v. Nat'l Australia Bank Ltd., 561 U.S. 247, 255, 261 (2010).

[390] Morrison v. Nat'l Australia Bank Ltd., 561 U.S. 247, 255 (2010) (citing Smith v. United States, 507 U.S. 197, 204, n. 5 (1993)).

[391] F. Hoffmann-La Roche Ltd. v. Empagran S. A., 542 U.S. 155, 164 (2004).

[392] *See supra* note 9.

may relocate their activities somewhere else with less stringent rules.[393] If more countries decide to follow similar approaches out of copyright concerns by having domestic exceptions and limitations apply to foreign training activities for AI models, this could quickly turn into a copyright nightmare for developers looking to launch an international AI product or service. Similarly, downstream AI developers and providers who are using an open-source model such as an LLM that is fine-tuned may also become liable for copyright infringement if the training data is encoded into the model parameters as discussed.[394] Without international harmonization, different regulators will inevitably come to strike their own balance between access to copyrighted works and protection of rightholders' interests, as they already have. Indeed, and as discussed,[395] copyright exceptions and limitations for text and data mining differ considerably between jurisdictions. This divergence becomes particularly challenging if the rules are each applied extraterritorially to the same underlying process. A model is typically only developed once due to the vast amounts of time and resources required. It may be infeasible to expect developers to deploy different datasets and training exercises for different markets subject to their own unique requirements.

In the best-case scenario, developers may decide to use a uniform training dataset and resulting model based on the most stringent requirements from larger markets. That would be similar to the Brussels effect but not necessarily centered on EU law this time around. The downside is that this would result in a potentially less capable model becoming standard in all countries if fewer pieces of data are being used due to copyright restrictions. If developers run into data scarcity issues as a consequence, some types of models or specific use cases might not be feasible to release at all. In the worst-case scenario, developers may decide to abandon certain markets entirely if the costs of reworking the development lifecycle outweigh the gains. Of course, excluding companies and individuals in selected countries from taking advantage of the latest AI developments would have very serious consequences for commerce, innovation, and technological progress. In the long-term, it could stifle competition for local industries, reduce the attractiveness

[393] In addition to the EU and Brazil, South Korea recently also proposed comprehensive AI regulations, the Korean AI Basic Act, which are applied extraterritorially to "any acts conducted abroad that affect the domestic market or users in the Republic of Korea." *See* Ingongjineung baljeon-gwa silroe giban joseong deung-e gwanhan gibonbeoban [Basic Act on the Development of Artificial Intelligence and the Establishment of Trust], Bill No. 2206772, art. 4, 22nd Nat'l Assembly (passed Dec. 26, 2024) (S. Kor.). Canada has introduced its draft Artificial Intelligence and Data Act in 2022, which, although pending review, appears to be drafted with the intent to give it extraterritorial reach should AI systems be designed, developed, or made available for use or managed in Canada. *See* Bill C-27, Artificial Intelligence and Data Act, pt. 3, 44th Parl., 1st Sess., 2022 (Can.). Countries that have previously followed the footsteps of the EU when adopting extraterritorial provisions governing data privacy are likely to do the same for AI. Canada was one of those countries, together with California, Singapore, Australia, and China, among others. *See* references *supra* note 13. It is still an open question, however, whether they will also apply their copyright laws extraterritorially for foreign training activities.

[394] *See supra* Section III.C.

[395] *See supra* Sections III.B, III.C.3.

of these markets for foreign enterprises, and impede the broader societal benefits that new AI technologies can offer.

4. The Tension Between Copyright Territoriality and the Extraterritorial Application of Copyright Laws

These extreme consequences compel the question whether an extraterritorial application of copyright laws for foreign training activities is compatible with the principle of territoriality and established international copyright norms. The exclusive rights afforded to rightholders are territoriality distinct in each country and can only be extended to conduct occurring domestically.[396] But that does not mean, as explained above,[397] that rightholders are without recourse when conduct is occurring across borders. Indeed, courts have been attentive to the risk of avoiding intellectual property laws in these circumstances by finding infringement where there is some form of territorial connection to the regulated conduct whether that is based on subjective or objective territoriality conditions.[398] This practice does not conflict with public international law. It is well-recognized that states are deemed to have a substantial margin of appreciation in stipulating the conditions for the applicability of national law, including in deciding what territorial factors are decisive for their laws to apply.[399] In the *S.S. Lotus* case, the Permanent Court of International Justice famously held that international jurisdiction is inherently territorial, and that a state may not exercise its power within the territory of another state.[400] However, the Court also concluded that did not prohibit states from exercising prescriptive jurisdiction within its own territory regarding acts which have taken place abroad.[401]

That states have a wide discretion in establishing what are the territorial connections for regulating conduct does not mean that there are no absolute limits or restraints. Public international law doctrines of non-intervention and sovereignty, which are equally shared between states, define the outer boundaries for what states are permitted to do.[402] Where states have concurrent jurisdiction over the same event, basic principles of international comity suggest that states are expected to exercise jurisdiction fairly and in good faith as a form of voluntary restraint.[403]

[396] *See supra* Section IV.

[397] *Id.*

[398] *Id.*

[399] *See* Barcelona Traction, Light and Power Co., Ltd. (Belgium v. Spain), Judgment, 1970 I.C.J. Rep 3, ¶ 70 (Feb. 5) (separate opinion by Fitzmaurice, J) (holding that "international law does not impose hard and fast rules on States delimiting spheres of national jurisdiction . . . but leaves to States a wide discretion in the matter").

[400] S.S. Lotus (Fr. v. Turk.), Judgment, 1927 P.C.I.J. (ser. A) No. 10, at 18-19 (Sept. 7).

[401] *Id.* at 19.

[402] Ryngaert, *supra* note 296, at 40; Mann, *supra* note 297, at 20-21; Maier, *supra* note 296, at 64-68.

[403] Maier, *supra* note 296, at 70-72; Harold G Maier, *Extraterritorial Jurisdiction at a Crossroads: An Intersection Between Public and Private International Law*, (1982) 76 AM. J. INT'L L. 280, 295-96 (1982).

Where there is a prescriptive conflict between states, the relative impact of regulating the event should be considered.[404] Yet there is far from international consensus on how to more precisely balance these competing contacts and, in turn, the interests that they represent.[405] There is no penalty for the failure to adequately consider foreign contacts and interests when prescribing and adjudicating foreign conduct. There is no supranational authority stopping regulators or courts from doing so. Practically speaking, the most likely result from an expansive extraterritorial approach would be that foreign courts refuse to recognize such judgments[406] or that diplomatic relationships between countries feel an impact.

There are two aspects complicating the present situation for copyright law. First, the EU AI Act makes no distinction between foreign and domestic training activities when regulating them from a copyright perspective. The message is clear: any AI model, regardless of where it was developed and where the data came from, must comply with EU copyright law, even if that is done retroactively. This is problematic for developers for the reasons set out above, but it is also legally flawed.

[404] Rättzén, *supra* note 285, at 405-406. Similar factors are included in the RESTATEMENT (THIRD) OF FOREIGN RELATIONS LAW § 403(2)(c) (Am. L. Inst. 1987) (considering when exercising prescriptive jurisdiction "the extent to which other states regulate such activities, and the degree to which the desirability of such regulation is generally accepted.").

[405] *See* David J. Gerber, *The Extraterritorial Application of German Antitrust Law*, 77 AM. J. INT'L L. 756 (1983) (stating that the international community has failed to develop jurisdictional principles accommodating both the needs of regulating states while avoiding impinging on the legitimate interests of other states); Kevin R. Roberts, *Extraterritorial Application of United States Antitrust Laws: Minimizing the Conflicts*, 1 U. MIAMI INT'L & COMP. L. REV. 325, 348 (2015) (stating that international rules on balancing sovereign interests are not fully developed); Professor Meeseen has proposed, although in the context of antitrust law and drawing from German experiences, that "a state is prohibited from taking measures of antitrust law if the regulatory interests it is pursuing are outweighed by the interests of one or more foreign states likely to be seriously injured by those measures." *See* Karl M. Meessen, *Antitrust Jurisdiction under Customary International Law*, 78 AM. J. INT'L L. 783, 804 (1984). Professors Grossfeld and Rogers have instead argued that deference to foreign mandatory law may be appropriate if that law "expresses values shared in common and which the receiving country is itself willing to protect." *See* Bernhard Grossfeld & C. Paul Rogers, *A Shared Values Approach to Jurisdictional Conflicts in International Economic Law*, 32 INT'L & COMP. L.Q. 931, 939 (1983). Frederick A. Mann, on the other hand, has rejected the notion considering and weighing foreign interests as a form of prescriptive restraint, suggesting that it is "nothing but a political consideration" and that it is "the objective test of the closeness of connection, of a sufficiently weighty point of contact between the facts and their legal assessment that is relevant. The lawyer balances contacts rather than interests." *See* Mann, *supra* note 297, at 30-31.

[406] Refusals to recognize foreign judgments are not uncommon. A well-known example is the Yahoo! litigation, where Yahoo! was sued in France to take down web pages posting the sale of Nazi memorabilia which were illegal in France. This resulted in an injunction, even if that post was accessible and available everywhere. *See* Tribunal de grande instance [TGI] [ordinary court of original jurisdiction] Paris, May 22, 2000, Interim Order No. 00/05308 (Fr.); The United States District Court of California subsequently refused to enforce the decision on the ground that it violated the First Amendment. *See* Yahoo!, Inc. v. La Ligue Contre Le Racisme et L'Antisemitisme, 169 F. Supp. 2d 1181, 1192 (N.D. Cal. 2001). *See also* Sarl Louis Feraud Int'l v. Viewfinder Inc., 406 F. Supp. 2d 274, 281-85 (S.D.N.Y. 2005) (refusing to enforce a French decision to award damages and injunctive relief for foreign copyright infringement on the basis of the First Amendment).

Foreign copying can, of course, only infringe foreign copyright. Then, there will be no territorial connection that can be localized in relation to the regulated conduct that is the act of reproduction. This is true notwithstanding the fact that the rules are designed to only apply when completed AI models are placed on the domestic or regional market. Copyright laws are meant to regulate the exercise of exclusive rights within their respective territories; they are not meant to regulate what exclusive rights exist in other countries, conditionally upon the domestic market that are *also* accessed at a later point in time. While states have a legitimate interest in regulating products and services placed on their domestic markets, other states arguably have a substantial interest in regulating the training activities ongoing in their own territory. Furthermore, a model that is developed in one country may be released in multiple countries that will each have a governing interest. If we allow conflicting copyright laws to extraterritorially and retroactively extend to entirely foreign conduct, we are creating an incredibly complex situation of concurrent jurisdiction. Although public international law permits the exercise of concurrent jurisdiction,[407] this assumes that states are regulating cross-border conduct[408] and not entirely foreign conduct for which there is no territorial connection.

Second, Article 5(2) of the Berne Convention provides that the extent of copyright protection shall be governed "exclusively" by the laws of the country where protection is claimed. That is to say, copyright laws can *only* regulate the exclusive rights, and the exceptions and limitations that apply to them, within their own territory.[409] This stems from the fact these rights legally only exist in their respective country of protection as a consequence of copyright territoriality. If copyright laws are employed to target foreign copying, which as such has no connecting factors to the home territory, then the extent of copyright protection in those foreign countries is no longer "exclusively" governed by the laws of the country where protection is claimed as required by the Berne Convention. Indeed, as the Ninth Circuit explained in *Subafilms*:

> We think it inappropriate for the courts to act in a manner that might disrupt Congress's efforts to secure a more stable international intellectual property regime unless Congress otherwise clearly has expressed its intent. The application of American copyright law to acts of infringement that occur *entirely* overseas clearly could have this effect. Extraterritorial application of American law would be contrary to the spirit of the Berne Convention and might offend other member nations by effectively displacing their law in circumstances in which previously it was assumed to govern.[410]

The only way to circumvent this problem would be to argue, as discussed above, that the obligation to comply extraterritorially with copyright law in the EU AI Act

[407] Laker Airways Ltd. v. Sabena, Belgian World Airlines, 731 F.2d 909, 952 (D.C. Cir. 1984) ("There is no principle of international law which abolishes concurrent jurisdiction.").

[408] *See supra*, Section IV (discussing how the exercise of legislative jurisdiction is premised on finding a sufficient territorial connection).

[409] *Id.*

[410] Subafilms, Ltd. v. MGM-Pathe Commc'n Co., 24 F.3d 1088, 1097 (9th Cir. 1994) (emphasis added).

is not concerned with substantive copyright law as such. The argument would boil down to that, because these provisions are separately formulated in AI regulations instead of copyright laws, it is the AI regulations and not substantive copyright law that have been extended extraterritorially. As discussed, it is also possible to interpret Recital 106 and Article 53.1(c) of the EU AI Act as merely requiring providers of general-purpose AI models to "put in place a policy" to comply with rightholders' opt-outs and that the consequence of the failure to do that is non-compliance of the EU AI Act, not copyright infringement.[411] The Berne Convention, or copyright territoriality more generally, has nothing to say about AI regulations. Yet this becomes a seemingly artificial distinction if the provisions are referring to substantive copyright law or are mirroring substantive copyright law standards, which indeed is the case here. Moreover, even if the remedy for failure to comply is not copyright infringement, but instead non-compliance of AI regulations, it could be argued that copyright territoriality as such is designed to maintain the sovereign independence of states, not rightholders. States, not rightholders, are the ultimate beneficiaries of copyright law that then grant such rights to rightholders through legislation. So, if AI regulations govern the same behavior and activity as copyright law but attach their own remedies for any such violations, then has foreign sovereignty not been impacted just as much? Clearly, the line to walk on is a very fine one, and there is considerable risk that the EU AI Act is incompatible with copyright territoriality unless there is a permissive rule in international law tolerating this.

The risk of such violations of international law is even greater for the Draft Brazilian AI Regulations. Article 63 of the draft regulations expressly excludes research organizations and institutions, among other named actors, from committing copyright infringement in certain circumstances when carried out text and data mining. This provision is extraterritorially and retroactively applied to foreign training activities due to Article 4 of the same draft regulations.[412] In addition, the draft regulations provide for administrative penalties if any of the rules are not followed.[413] This appears to suggest that failing to comply with the requirements for the copyright exception for text and data mining in Article 63 will result in *both* copyright infringement and non-compliance with the Draft Brazilian AI Regulations. If this interpretation is correct, then the Brazilian rules will likely conflict with copyright territoriality as enshrined in Article 5(2) of the Berne Convention, although in a limited fashion with respect to research organizations and institutions. It is unclear if the obligation for AI agents (including "developers," "distributors" and "applicators") to remunerate rightholders under Article 65 of the draft regulations will also be applied extraterritorially for use of protected content in "processes of mining, training or development of AI systems." To the extent that

[411] *See supra* Section VII.A.I.

[412] As discussed, Article 4(V) extends the scope of the entire draft AI regulations, including Article 63, to anyone developing an AI system with the intention of placing it on the Brazilian market or applying it in services to Brazilian users. *See* Draft Brazilian AI Regulations, *supra* note 379.

[413] *Id.* at art. 50.

Article 65 is applied extraterritorially, then this would arguably collide with foreign sovereignty and be incompatible with Article 5(2) of the Berne Convention. This is because there can be no right to remuneration under copyright law without infringement, and there can be no infringement without an infringing act occurring in the relevant country of protection, under the principle of territoriality. Therefore, if commercial AI developers carry out the text and data mining activities outside Brazil, then the allegedly infringing act of reproduction would not take place in Brazil. Brazilian copyright law should, arguably, never be applied under such circumstances.

B. Rethinking Model Training as a Product-By-Process Problem

The main issue with applying existing copyright laws or AI regulations extraterritorially to target foreign training activities is that the conduct is exclusively occurring outside the country of protection. This is problematic because the exclusive rights relevant for the acts of reproduction are those of the foreign country, where the model training occurs, not where the completed model is eventually released. From the rightholders' point of view, this distinction makes little sense. What rightholders worry most about is the completed AI model and if that AI model is subsequently marketed and sold to customers in their home market. It is the use, not the underlying creation, of the generative AI model that could harm rightholders by reducing demand for their original works.[414] Yet, and as discussed,[415] the vast majority of AI-generated works will not directly infringe the training data. They will often not be infringing derivative works or will not be identical or substantially similar to the training data. This obviously makes the situation for rightholders much more difficult. Copyright statutes are formulated to regulate copying, which is the reproduction of works and the dissemination of those works to the public. Copyright law is not designed to deal with new works and therefore, new reproductions which are statistically created from a process using vast quantities of original works in fragmented parts. This requires us to critically rethink how copyright law is framed.

The legal dilemma is similar to what used to be the case for products derived from processes in patent law. A method or process patent protects the carrying out of that method or process. Patent laws also commonly provide exclusive rights for the sale or use of products derived from patented processes.[416] Because patents are territorial rights, like copyright, they cannot have an extraterritorial effect unless

[414] *See supra* Section III.B.1 (discussing, in relation to fair use, how AI-generated works can reduce the demand of original authors' works). *See also infra* Section VII.C and note 440 (discussing how these arguments have been framed by plaintiffs in recent copyright complaints).

[415] *See supra* Section III.B.1 (discussing idea-expression dichotomy in relation to the fair use test). *See also infra* Section VIII.A (discussing how AI-generated works will only exceptionally be infringing derivative works).

[416] G.H.C. BODENHAUSEN, GUIDE TO THE APPLICATION OF THE PARIS CONVENTION FOR THE PROTECTION OF INDUSTRIAL PROPERTY 85 (1967).

expressly provided for in statute.[417] This meant that, if someone would carry out the protected method or process of a patent, which was registered and granted in country A, in another country B, and obtained products as a result of that method or process, there could be no patent infringement in country A. This would be the case even if those derived products are subsequently imported to country A. This is because the exclusive rights afforded to the patent holder for the method or process patent would only extend to the act of using the patented method or process, or the acts of selling or using products derived thereof, if those acts took place within a country where the patent was registered.[418] This was particularly problematic because patentees commonly only apply for patent protection in a few countries, meaning that patent infringement could easily be avoided for method and process claims by relocating the infringing activities to another country where there was no granted patent. What is more, economic value for the use of a method or process patent typically also lies in the product derived thereof, not the carrying out of the method or process.[419]

Patent laws have since been amended in many countries to close this loophole. Now, in jurisdictions such as the United States,[420] U.K.,[421] and Europe,[422] method

[417] *See* H.R. REP. NO. 99-807, at 5 (1986) ("American patent law — like the law of other nations — does not have an extraterritorial effect. To provide that American law should govern conduct that occurs in other countries would conflict with basic notions of national sovereignty. For that reason, American patent law has always required that the infringing act occur within the United States territory."). For a summary about the territoriality of patents, *see* Rättzén, *supra* note 285, at 362-65.

[418] *See* H.R. REP. NO. 99-807, at 5 (1986) ("With respect to process patents, courts have reasoned that the only act of infringement is the act of making through the use of a patented process; therefore, there can be no infringement if that act occurs outside the United States."). *See also* Timothy R. Holbrook, *Extraterritoriality in U.S. Patent Law*, 49 WM. & MARY L. REV. 2119, 2139 (2008); Dan L. Burk, *Patents in Cyberspace: Territoriality and Infringement on Global Computer Networks*, 68 TUL. L. REV. 1, 47 (1993).

[419] *See* AMIRAM BENYAMINI, PATENT INFRINGEMENT IN THE EUROPEAN COMMUNITY 157-58 (1993).

[420] 35 U.S.C. § 271(g) (2021) ("Whoever without authority imports into the United States or offers to sell, sells, or uses within the United States a product which is made by a process patented in the United States shall be liable as an infringer, if the importation, offer to sell, sale, or use of the product occurs during the term of such process patent.").

[421] Patents Act 1977, § 60(1)(c) (U.K.) ("[I]f he does any of the following things in the United Kingdom in relation to the invention without the consent of the proprietor of the patent . . . where the invention is a process, he disposes of, offers to dispose of, uses or imports any product obtained directly by means of that process or keeps any such product whether for disposal or otherwise.").

[422] Agreement on a Unified Patent Court, art. 25(c), 2013 O.J. (C 175) 1, 8 ("offering, placing on the market, using, or importing or storing for those purposes a product obtained directly by a process which is the subject-matter of the patent"). Before the introduction of the unified patent, which is a European unitary right, national patent laws of European countries still had similar provisions in respect of national patents. These provisions were based on the Agreement relating to Community Patents, art. 25(c), 1989 O.J. (L 401) 1 (EC) [hereinafter CPC] ("A Community patent shall confer on its proprietor the right to prevent all third parties not having his consent . . . from offering, putting on the market, using, or importing or stocking for these purposes the product obtained directly by a process which is the subject-matter of the patent."). The CPC was signed by twelve Member States, it was never ratified by enough Member States to enter into force. However,

or process patents protect not only the carrying out of the method or process but also against the importation of products derived from such methods or processes carried out abroad.[423] Therefore, even if there would be no infringement in the country where the patented method or process was carried out in its entirety, patent laws are now applied extraterritorially to regulate such offshore conduct that can be traced to derivative products. A House report before the U.S. Congress in 1986, when it was considered whether such a provision should be introduced into the U.S. Patents Act, explained as follows: "[w]ithout domestic legal protection, competitors using the protected process may accept the limited risks of foreign production and importation, in exchange for lower foreign production costs. There is no policy justification for encouraging such overseas production and concurrent violation of United States intellectual property rights."[424]

The same logic might be applied to the present case of model training and copyright law. Model training and copyright infringement could also be thought of as a product-by-process problem. Similarly, as in the case of products derived from patented processes in patent law, what is most economically threatening to rightholders in copyright law is the generated output, i.e., the resulting products from generative AI models. The fact that copyrighted materials themselves were used in the underlying development and training process is arguably of less value, as it does not directly impact the end market or value. It is the generated output, or the ability to generate output, that could reduce demand from rightholders' original works and potentially harm their ability to make a living as authors.[425] However, and as discussed, the generated output will not directly infringe on copyright in the vast majority of cases. This means that rightholders' only legal option is to argue that the underlying development process of using their copyrighted works as training data infringes the exclusive right to reproduce their works. As discussed in this Article, using copyrighted materials for text and data mining purposes is a process that may infringe on copyright in some countries but not in others.[426] This creates a transnational data loophole that can be exploited, which is similar to what previously existed in patent law for products derived from processes. Model training activities can be offshored to AI-friendly jurisdictions, and once the model is complete, it can be sold and marketed everywhere else in the world, including in countries where the same training activity would be infringing.[427]

many European countries have still included its provisions, including Article 25(c) of the CPC, into their national patent laws.

[423] The same right is also provided for in the TRIPS Agreement, *supra* note 130, at art. 28(1)(b).

[424] H.R. REP. NO. 99-807, at 6 (1986). *See also* Amgen Inc. v. U.S. Intern. Trade Com'n, 902 F.2d 1532, 1538-39 (Fed. Cir. 1990); Bayer AG v. Housey Pharms., Inc., 340 F.3d 1367, 1375 (Fed. Cir. 2003).

[425] *See supra* Section III.B.1 (discussing, in relation to fair use, how AI-generated works can reduce the demand of original authors' works). *See also infra* Section VII.C, note 440 (discussing how these arguments have been framed by plaintiffs in recent copyright complaints).

[426] *See supra* Section III.B.

[427] *See supra* Section VI.

The similarities between these problems in patent law and copyright law suggest that we should consider adopting a similar legislative approach. If policymakers consider that rightholders should have legal recourse to target foreign acts of reproduction for the purpose of text and data mining then introducing a new exclusive right could be more suitable as opposed to carving out extraterritorial exceptions and limitations. This would be preferable due to the significant problems with a far-reaching extraterritoriality approach, as discussed above.[428] An example for a statutory provision could read as follows:

> *1. Whoever places an AI system or model on the market within [home country] shall be liable for copyright infringement if the development or deployment of such system or model involved the automated use of works in data and text mining related processes, without authorization from rightholders where such works are protected by copyright in [home country], irrespective where such automated use took place.*
>
> *2. The following acts shall be exempted from liability provided for in paragraph 1 above:[details of exempted acts depending on national policy preferences]*

The first paragraph would establish a new exclusive right based on conduct that *actually* occurs where copyright protection is claimed, that is the placing of the AI system or model on the domestic market, regardless of where the reproduction of training data occurs. Such a provision would be more aligned with copyright territoriality but will still indirectly regulate foreign copying activities. The second paragraph would allow states to set out any exceptions and limitations they deem fit for this new exclusive right. Of course, if a country considers that text and data mining should always be permitted and never infringe on copyright, then this type of provision is not necessary.

However, even if such a provision is introduced in national copyright laws, it is not straightforward whether the same logic from patent law can be directly extended to copyright law and model training. While it is a product-by-process problem we are similarly dealing with, there is far from international consensus from a copyright perspective for regulating data collection, storing, and processing required for training AI models.[429] Regulations and their interpretations by courts are rapidly evolving across the world and there is a high likelihood that the same type of training activity will be infringing in some countries and not infringing in others. The situation for product-by-process patents is different in that regard. No one would seriously dispute that infringing a patent should be condemned, *if* a valid patent exists in that country. Moreover, copyright protection is universally automatic across the world without any formalities[430] in contrast to patent protection, which must be applied for and registered in each country.

[428] *See supra* Section VII.A.4.

[429] *See supra* Section III.B.

[430] Berne Convention, *supra* note 92, at art. 5(2) (stating that "[t]he enjoyment and the exercise of these rights shall not be subject to any formality").

What is more, international patent treaties do not impose the same territoriality restriction as the last sentence of Article 5(2) of the Berne Convention. But with or without any such exclusivity restriction, it is clear that all intellectual property rights are inherently territorial.[431] Intellectual property rights are national rights, copyright and patents alike, and therefore are always exclusively governed by the national laws responsible for creating them. There may simply be no need to state something so obvious in all international intellectual property treaties. However, what arguably makes the situation for product-by-process patents legally unique is that there is a basis for including these types of extraterritorial import provisions in international law. Article 5 *quater* of the Paris Convention[432] sets out that "[w]hen a product is imported into a country of the Union where there exists a patent protecting a process of manufacture of the said product, the patentee shall have all the rights, with regard to the imported product, that are accorded to him by the legislation of the country of importation, on the basis of the process patent, with respect to products manufactured in that country." Although ambiguously worded, this means that if a country's patent laws provide protection for products derived from patented processes, then that country's laws will also apply to products which are imported into that country.[433] This applies even if the patent laws of the country where the product was manufactured using the patented process did not provide such protection.[434] For example, if someone uses a patented process to obtain products derived thereof in country A, where there is no granted patent or where patent law does not provide protection for products derived from patented processes, and later imports those products into country B, where there is a granted patent and where such protection is provided in patent law, then country B's laws will also apply to the imported product. In this way, Article 5 *quater* does not harmonize product-by-process patent laws but permits the extraterritorial application of national patent law to imported products derived from foreign processes to the extent that a particular country has introduced such protection in their own patent statute.

The extraterritorial reach of patent laws for product-by-process patents is therefore explicitly mandated in international patent law. No such provision can be found in any international copyright treaties. If regulators are to extraterritorially extend their copyright laws to exclusively foreign conduct, whether it is done directly as is the case in the EU AI Act, or indirectly as in the case of a new exclusive right for placing AI systems or models on the market, then a similar provision may be required to preserve foreign sovereignty as a matter of international copyright law. Indeed, it could be argued that the absence of any such mandate when read together

[431] *See*, *e.g.* Lundstedt, *supra* note 295, at 91-104 (summarizing the territoriality limitations of intellectual property rights).

[432] Paris Convention for the Protection of Industrial Property, Mar. 20, 1883, as revised in Stolkholm, July 14, 1967, 21 U.S.T. 1583, 24 U.S.T. 2140 [hereinafter Paris Convention].

[433] BODENHAUSEN, *supra* note 416, at 85; Regina A. Loughran, *The United States Position on Revising the Paris Convention: Quid Pro Quo or Denunciation*, 5 FORDHAM INT'L L.J. 411, 430 (1982).

[434] Loughran, *supra* note 433, at 430.

with Article 5(2) of the Berne Convention would by default prohibit states from regulating foreign copying activities, in any form.

C. *Why Should We Bother with Closing the Transnational Data Loophole?*

It is worth pausing to reflect on why we should at all bother with closing the transnational data loophole. In principle, the transnational data loophole as it currently stands allows AI developers to exploit the fact that copyright laws are territorial in nature and that there are more AI-friendly jurisdictions that clearly provide exceptions and limitations for text and data mining. Once the AI model is complete after pre-training, it can be sold and marketed anywhere else in the world, even in countries where that activity would have been infringing *if* it occurred there.

The EU expressly considered this to be problematic in the EU AI Act when it extraterritorially extended the obligation of respecting opt-outs under Article 4(3) of the Copyright Directive for third country providers of general-purpose AI models. In particular, as stated in Recital 106, this extraterritorial application was necessary "to ensure a level playing field among providers of general-purpose AI models where no provider should be able to gain a competitive advantage in the Union market by applying lower copyright standards than those provided in the Union." The expressed concern, therefore, was that models that have been developed based on diverging copyright requirements could end up competing on the same market. Because access to data is fundamental to the functioning of AI models, this could mean that models developed in other countries under more lax copyright rules would obtain a competitive advantage. Although this makes sense, it arguably fails to consider the interests of rightholders. This Article has explained that the offshoring training activities to foreign countries, where text and data mining is clearly a permitted activity, would make the copyright protection worthless for the rightholders in those countries where this is not the case.[435] This critical point is completely missing from the EU narrative and it is unclear as to why.

Still, it is questionable whether we are putting too much emphasis on the exclusive right of reproduction for model training and exaggerating its commercial significance. What is it that we hope to achieve? If a model has been trained on one billion copyrighted works without the permission of their rightholders, each rightholder would have an extremely diluted commercial stake in the total model, where each work is only a statistical representation among many millions of other parameters. Statutory damages may be the only viable remedy in such cases, in the countries that have them, such as the United States.[436] In other countries that do not have statutory damages, only nominal damages may be available for individual rightholders.[437] That is not to say that total damages awards cannot be high in these circumstances. Collectively, the rightholders may have a strong, joint commercial

[435] *See supra* Section VI.

[436] 17 U.S.C. § 504(c)(1).

[437] *See* cases cited *supra* note 261.

interest in the complete AI model, but individually it will be very minimal, if even possible, to meaningfully measure without the help of statutory tools.

There is a rapidly growing marketplace for training data, illustrating that model training is indeed a commercially viable practice of exploiting copyrighted works.[438] Many of the copyright complaints which have been filed against AI developers and providers therefore request compensation for the loss of commission or license fee that rightholders have endured by the use of their copyrighted materials without permission for text and data mining.[439] However, when reviewing these complaints more closely, it is clear that what many rightholders are primarily concerned with is how AI-generated works compete in the same marketplace as the original copyrighted works, thereby reducing the demand for such works, ultimately threatening their livelihood and/or their ability to control how their works are used.[440] The end use of the generative AI model

[438] *See supra* Section III.B.1 (referring to, for example, OpenAI recently striking various licensing deals to obtain more training data, discussed in relation to the fourth fair use factor).

[439] *See, e.g.,* Complaint at 4, Authors Guild v. OpenAI Inc., No. 1:23-cv-08292 (S.D.N.Y. filed Sept. 19, 2023) (alleging that "[d]efendants could have 'trained' their LLMs on works in the public domain. They could have paid a reasonable licensing fee to use copyrighted works" and requesting award of statutory damages up to $150,000 per infringed work to the plaintiffs and class members); Complaint at 90, *Daily News v. Microsoft*, No. 1:24-cv-03285 (S.D.N.Y. filed Apr 30, 2024) (alleging that "[d]efendants repeatedly copied the Publishers' Works, without any license or other compensation to the Publishers" and requesting statutory damages, actual damages and restitution of profits); Complaint at 56, *Getty Images v. Stability AI*, No. 1:23-cv-00135 (D. Del. filed Feb 3, 2023) (alleging that "[w]hile Getty Images licenses its proprietary content to responsible actors in appropriate circumstances, Stability AI has taken that same content from Getty Images without permission, depriving Getty Images and its contributors of fair compensation, and without providing adequate protections for the privacy and dignity interests of individuals depicted" and requesting damages and any additional profits, or alternatively statutory damages of up to $150,000 for each infringed work); Complaint at 92, N.Y. Times v. Microsoft Corp., No. 23-cv-11195 (S.D.N.Y. filed Dec. 27, 2023) (alleging that "[d]efendants repeatedly copied this mass of Times copyrighted content, without any license or other compensation to The Times" and requesting statutory damages, actual damages and restitution of profits).

[440] *See, e.g.*, Complaint at 5, Andersen v. Stability AI Ltd., No. 23-cv-00201-WHO (N.D. Cal. Filed Oct. 30, 2023) ("These resulting derived images compete in the marketplace with the original images. Until now, when a purchaser seeks a new image 'in the style' of a given artist, they must pay to commission or license an original image from that artist. Now, those purchasers can use the artist's works contained in Stable Diffusion along with the artist's name to generate new works in the artist's style without compensating the artist at all."); Complaint at 2, Authors Guild v. OpenAI Inc., No. 1:23-cv-08292 (S.D.N.Y. filed Sept. 19, 2023) ("Defendants' LLMs endanger fiction writers' ability to make a living, in that the LLMs allow anyone to generate—automatically and freely (or very cheaply)—texts that they would otherwise pay writers to create. Moreover, Defendants' LLMs can spit out derivative works: material that is based on, mimics, summarizes, or paraphrases Plaintiffs' works, and harms the market for them"); Complaint at 7, Daily News v. Microsoft, No. 1:24-cv-03285 (S.D.N.Y. filed Apr 30, 2024) ("Defendants are taking the Publishers' work with impunity and are using the Publishers' journalism to create GenAI products that undermine the Publishers' core businesses by retransmitting 'their content'—in some cases verbatim from the Publishers' paywalled websites—to their readers"); Complaint at 9, Getty Images v. Stability AI, No. 1:23-cv-00135 (D. Del. filed Feb 3, 2023) ("Stability AI now competes directly with Getty Images by marketing Stable Diffusion and its DreamStudio interface to those seeking creative imagery"); Complaint at 2, N.Y. Times v. Microsoft Corp., No. 23-cv-11195 (S.D.N.Y.

presents a very different policy problem than controlling what training data is inputted for creating the model. It may be that jurisdictions will diverge in their policy priorities in this regard. Some states might consider both the unauthorized use of copyrighted materials as training data and the end use of generative AI models should be regulated from a copyright perspective. Other states might consider that only one of the two should be regulated. This policy discussion is highly relevant for the issue of the transnational data loophole, as this problem only arises from the unauthorized use of copyrighted materials as training data. If this is not what is of interest when regulating from a policy perspective, then policymakers should instead focus on what liability may attach to the distribution of AI-generated works themselves. This is discussed in the following section.

VIII. Avoiding the Transnational Data Loophole

A. Whether AI-Generated Works Are Infringing Derivative Works

It has been concluded above that it is possible that we cannot adequately close the transnational data loophole as the law currently stands without engaging in far-reaching extraterritoriality arguments. An extraterritorial application of copyright law to what is entirely foreign conduct could unreasonably interfere with foreign sovereignty unless international copyright treaties are revised accordingly. It may also be the case that some states consider that, from a policy perspective, what is problematic is not necessarily the unauthorized reproduction of copyrighted materials for the purpose of text and data mining, but how the completed AI model is being used on the market for generating synthetic works. After all, it is the AI-generated works, not the underlying training data, which could interfere with rightholders' ability to commercialize their own, original works.

This raises the question whether the focus of trying to close the transnational data loophole associated with model training is incorrectly framed, and whether we should focus on regulating the outputted AI-generated content instead of the inputted training data. Indeed, some of the pending lawsuits against developers submit that the generated output from generative AI models is derivative of authors' original works. The New York Times alleged in its lawsuit against OpenAI that "memorized examples constitute unauthorized copies or derivative works of the Times Works used to train the model."[441] Similarly, some plaintiffs alleged that Stable Diffusion's model could produce images that are "highly similar to or derivative of the Getty Images proprietary content that Stability AI copied extensively in the course of training the model."[442] This point is important, because if generated output is derivative of the original works, then rightholders have recourse against new primary infringing acts. Crucially, these infringing acts would

filed Dec. 27, 2023) ("Defendants' unlawful use of The Times's work to create artificial intelligence products that compete with it threatens The Times's ability to provide that service").

[441] Complaint at 98, N.Y. Times v. Microsoft Corp., No. 23-cv-11195 (S.D.N.Y. filed Dec. 27, 2023).

[442] Complaint at 8, 61, Getty Images v. Stability AI, No. 1:23-cv-00135 (D. Del. Filed Feb 3, 2023).

take place, at least, where the end user prompting the model is located. This would mean that any extraterritoriality issues related to where the model training occurs become a moot point. Therefore, instead of closing any transnational data loophole, we are now refreshingly avoiding the problem altogether.

Although derivative works can be protected themselves by copyright, that is without prejudice to the copyright in the original work.[443] A license must therefore be obtained from the original rightholders if intending to create and commercially exploit derivative works.[444] Derivative works are works that incorporate portions of an original work but build upon that by recasting, transforming, or adapting it.[445] The amount of change needed for a derivative work to be protected is not always clear and differs between jurisdictions.[446]

[443] *See* Berne Convention, *supra* note 92, at art. 2(3) ("Translations, adaptations, arrangements of music and other alterations of a literary or artistic work shall be protected as original works without prejudice to the copyright in the original work").

[444] Russell v. Price, 448 F. Supp. 303, 304 (C.D. Cal. 1977), aff'd, 612 F.2d 1123 (9th Cir. 1979), cert. denied, 446 U.S. 952 (1980) ("[i]f one intends to create and commercially exploit a work that is derivative of a copyrighted work, a license to do so must be obtained from the owner.").

[445] *See* 17 U.S.C. § 101. Copyright statutes outside the United States also recognize that adaptations of copyrighted works, such as translations or alterations, can constitute works themselves. Although not codified in U.K. or European copyright law, similar protection is offered in national law for adaptations as is the case in the United States for derivative works. *See also* Copyright, Designs and Patents Act 1988, c. 48, § 21(3) (U.K.) (defining adaptations as a translation of the literary work, excluding computer programs and databases, a version of a dramatic work or a version of a literary work in which the story or action is conveyed wholly or mainly by means of pictures in a form suitable for reproduction in a book, or in a newspaper, magazine or similar periodical); Urheberrechtsgesetz [UrhG] [German Copyright Act], June 23, 2001, BUNDESGESETZBLATT, TEIL I [BGBL I] at 1858, § 3 (Ger.) ("Translations and other adaptations of a work which are the adapter's own intellectual creations are protected as independent works without prejudice to the copyright in the adapted work"); CODE DE LA PROPRIÉTÉ INTELLECTUELLE [C. Intell. Prop.] [Intellectual Property Code] arts. L112-3, L113 (Fr.) ("The authors of translations, adaptations, transformations or arrangements of works of the mind shall enjoy the protection afforded by this Code, without prejudice to the rights of the author of the original work;" "'[c]omposite work' shall mean a new work in which a preexisting work is incorporated without the collaboration of the author of the latter work;" "[a] composite work shall be the property of the author who has produced it, subject to the rights of the author of the preexisting work.").

[446] *See, e.g.*, Caffey v. Cook, 409 F. Supp. 2d 484, 496 (S.D.N.Y. 2006) (the new work must have "substantially copied" the original work in order to be considered a derivative work); Gracen v. Bradford Exchange, 698 F.2d 300, 303 (7th Cir. 1983) (the derivative work must be "substantially different" than the base work to be protected); Eden Toys v. Florelee Undergarment Co., 697 F.2d 27, 34-35 (2d Cir. 1982) (holding that smoothing the lines of a sketch to give it a "cleaner look" was sufficient to become derivative, even though the new work did not have a "different aesthetic appeal" from the original work); Alfred Bell & Co. Ltd. v. Catalda Fine Arts, Inc., 191 F.2d 99, 103 (2d Cir. 1951) (holding that something more than a "mere trivial variation" was required, even if that could occur accidentally); Macmillan v. Cooper (1924) 40 TLR. 186, 188 (U.K.) (holding that the derivative work must be the product of a "material change" when compared with the pre-existing work); Interlego A.G. v. Tyco Indus. Inc. [1988] RPC 343, 371-72 (U.K.) (holding that there must "in addition be some element of material alteration or embellishment which suffices to make the totality of the work an original work," but also stating that even a "relatively small alteration or addition qualitatively" may, if it is material, convert it into a new, original work).

In the United States, the question of whether AI-generated works are derivative will ultimately turn on whether there is a sufficient non-trivial and expressive variation in the derivative work to make it distinguishable from the pre-existing work in some meaningful way.[447] The answer to that question is complicated by the fact that AI-generated output is a result of a multimodal process, where typically nothing has been added to pre-existing works as such. The machine learning algorithms find patterns in vast data quantities that are aligned for generating works that are statistically similar.[448] AI-generated content therefore does typically not directly "derive" from specific pre-existing works as training data but from statistical representations and variations of those works.[449] The original input from the user prompting the model may also range widely depending on the use case, with some being more imaginative and others only rudimentary. Both aspects complicate the analysis of whether AI-generated content may be derivative, making it highly unlikely to be valid in all circumstances. Some AI-generated content may be derivative, others may not.[450]

But even assuming that AI-generated works could be derivative works in certain cases, this does not win an infringement case. The fact that a work is a derivative work does not automatically mean it is infringing on the pre-existing work. In the United States, a similar but not identical test for substantial similarity applies to derivative works to assess infringement.[451] The relevant question is whether the creative choices that define the pre-existing work were copied in the derivative work, without fundamentally altering the original message.[452] If the derivative work does not add anything "new and different" to the pre-existing work, then it is more likely that the derivative work will infringe the pre-existing work.[453] The impact that the derivative work has on the market of the pre-existing work further serves as a useful proxy for answering that question, but should ultimately remain a secondary consideration.[454] No separate infringement test exists for derivative works in the U.K. or the EU, which instead asks whether the portion copied of the whole pre-existing work itself meets the originality threshold.[455] Thus, the question

[447] Schrock v. Learning Curve Int'l, Inc., 586 F.3d 513, 521 (7th Cir. 2009); L. Batlin & Son, Inc. v. Snyder, 536 F.2d 486, 491 (2d Cir. 1976).

[448] Daniel Gervais, *AI Derivatives: The Application to the Derivative Work Right to Literary and Artistic Production of AI Machines*, 52 STAN. L. REV. 1111, 1128-31 (2022).

[449] *Id.*

[450] *Id.*

[451] *Id.* at 1125-26.

[452] Daniel J. Gervais, *The Derivative Right, or Why Copyright Law Protects Foxes Better than Hedgehogs*, 15 VAND. J. ENT. & TECH. L. 785, 843 (2013).

[453] *See* Lee v. Deck the Walls, 925 F. Supp. 576, 580-82 (N.D. Ill. 1996), *aff'd sub nom.* Lee v. A.R.T. Co., 125 F.3d 580 (7th Cir. 1997).

[454] Gervais, *supra* note 452, at 844-46 (further citing, regarding estimating market impact(s) on pre-existing works, Lee v. A.R.T. Co., 125 F.3d 580 (7th Cir. 1997); Mirage Editions, Inc. v. Albuquerque A.R.T. Co., 856 F.2d 1341 (9th Cir. 1988); Mufioz v. Albuquerque A.R.T. Co., 829 F. Supp. 309 (D. Ala. 1993), *aff'd* 38 F.3d 1218 (9th Cir. 1994)).

[455] *See supra* Section III.A (referring to Case C-5/08, *Infopaq I* , 2009 E.C.R. I-06569, ¶¶ 47-51). This dual originality-infringement methodology is not without criticism. *See* Amy B. Cohen,

in the U.K. and the EU is whether something that "contains the expression of the author's own intellectual creation" has been copied.[456]

It remains an open question at this point whether AI-generated works will meet these infringement tests.[457] Some AI-generated works will more closely resemble original training data, particularly if the training data has been memorized.[458] But in the vast majority of cases, AI-generated works may merely copy the style of a particular group of works, or blend multiple different styles or different content. This will not be very different from a human author who has inspected one or several original works, became inspired by them, and tried to produce something new but drew from those sources. As discussed, the idea-expression dichotomy prevents copyright from extending to mere ideas, including genre and styles, which should be "common building blocks" that cannot be monopolized.[459] Copyright laws have been designed to protect authors from the copying of their original expressions, not from moderately similar content that is only inspired, in whole or in part, by their works.[460] Broadly extending the concept of specifically infringing derivative works to all AI-generated works could have unforeseen and potentially dangerous consequences in balancing the scope of copyright protection and run counter to the idea-expression dichotomy, which seeks to preserve that delicate balance. Courts should therefore tread very carefully.

Yet the fact that AI-generated works may not themselves infringe on the copyright of rightholders does not mean that rightholders are not negatively impaired by them, and it is perhaps this what is most controversial today from a policy standpoint. Generative AI completely changes how creative works are created,[461] and may reduce demand for works produced by original authors even if

When Does a Work Infringe the Derivative Works Right of a Copyright Owner?, 17 CARDOZO ARTS & ENT. L.J. 623, 646-48 (1999) (criticizing the conflation of originality and infringement tests).

[456] *Infopaq I*, 2009 E.C.R., ¶ 48; Rosati, *supra* note 106, at 71.

[457] *See* Gervais, *supra* note 452, at 1128-31 (discussing various hypothetical scenarios where AI-generated works may or may not be infringing pre-existing works as derivatives).

[458] *See supra* Section III.C.1 (discussing to what extent generative AI models encode or "memorize" training data into the model parameters, in which case they will also be more prone to generate such original training data in an infringing way upon prompting by the end user*).*

[459] *See supra* Section III.B.1 (discussing idea-expression dichotomy in relation to the fair use test).

[460] Gates Rubber Co. v. Bando Chemical Industries, Ltd., 9 F.3d 823, 838 (10th Cir. 1993) ("Under the scenes a faire doctrine, we deny protection to those expressions that are standard, stock, or common to a particular topic or that necessarily follow from a common theme or setting. . . . Granting copyright protection to the necessary incidents of an idea would effectively afford a monopoly to the first programmer to express those ideas."); *See also Infopaq I*, 2009 E.C.R., ¶ 48 (copying part of a work only amounts to infringement if that part "contains the expression of the author's own intellectual creation," meaning that other, non-original aspects of a work may be copied without infringing).

[461] *See* Mark Lemley, *How Generative AI Turns Copyright Upside Down*, 25 COLUM. SCI. & TECH. L. REV. 22 (2023) (discussing how prompt-based generative AI models challenge traditional copyright norms of both originality and infringement).

not directly infringing.[462] Copyright laws do not protect authors against the reduction of demand for their works and it is ultimately a question for policymakers whether we think this should be the case.

B. Whether Dataset Creators and Model Providers Could be Secondarily Liable for Copyright Infringement

A significant downside with arguing that the user-generated content is infringing itself as a derivative work is that users are oblivious to what training data has been used. Because end-users have no access to the original training data and cannot see what data has been used to render a particular image, let alone control it, they have no means of avoiding copyright infringement when producing AI-generated works using models from third party developers or providers. Because copyright infringement is a strict liability tort,[463] it does not matter whether or not the end-user knows that they are committing an act of infringement when producing a "memorized" or "nearly memorized" version of an original work when prompting the model, which could count as an infringing derivative work. If we are satisfied with the proposition that AI-generated works may exceptionally be derivative works which infringe on the copyright of original training data, then it is similar to roll the dice for the end-user when prompting the model. This is not acceptable from a policy perspective and suggests that we should consider other infringing acts if we are to avoid the transnational data loophole when focusing on the output AI-generated content rather than input training data.

If the issue from a policy perspective is that generative AI models can produce "memorized" versions of training data, which directly compete with the original works and harm rightholders, then we need to accept the fact that this is the exception, not the norm. The vast majority of AI-generated works will *not* be "memorized" versions of the original training data.[464] It would seem unduly strict to punish the masses for the few and raises the question whether there could be a *de minimis* defense in this regard for model providers. However, the issue is that copyright law *de minimis* defenses, which can be raised in the United States, are outdated and ill-suited for this type of situation. *De minimis* goes to the trivial amount of copying of one or several works, not to any *de minimis* use or display of such works.[465] The issue with AI-generated works which are "memorized" is that

[462] *See supra* Sections II.B.1 (discussing the possibility for AI-generated works to reduce demand for original works in relation to the fair use test), VII.C, (discussing "reduction of demand" related arguments in filed copyright complaints). *See also* Lee et al., *supra* note 7, at 107 ("The outputs of a non-generative AI do not compete in the market for a copyrighted work in the sense that the fourth factor cares about. It is possible that these outputs could reduce the demand for the copyrighted work.").

[463] *See supra* references in note 197.

[464] *See supra* Section III.C.1.

[465] Bell v. Wilmott Storage Servs., 12 F.4th 1065, 1076 (9th Cir. 2021) ("[w]e reiterated that the de minimis concept applies to the amount or substantiality of the copying—and not the extent of the defendant's use of the infringing work"); Warner Bros. Inc. v. American Broadcasting Companies, Inc., 720 F.2d 231, 242 (2d Cir. 1983); Louisiana Contractors Licensing Serv., Inc. v. Am. Contractors Exam Servs., Inc., 13 F. Supp. 3d 547, 554 (M.D. La. 2014), *aff'd*, 594 F. App'x

they fully produce infringing copies of original works, not merely small fragments of such works, so the generated work is therefore unlikely to be non-trivial. Another issue is that *de minimis* defenses against copyright infringement are not recognized in many other parts of the world, including the EU.[466]

The next best argument is that model providers are secondarily liable for any primary copyright infringement committed by the end-users of their model. Secondary liability has a significant advantage for rightholders from an extraterritoriality perspective. As discussed, the prevailing view is that the contributory act in a cross-border situation shall follow the law applicable to the primary infringing act.[467] This means that model providers cannot escape secondary copyright infringement by relocating their operations to another country. Even if a model provider is located abroad and supplies their generative AI model to users residing in, for example the United States, the provider will become secondarily liable for any primary infringement committed by those users.[468]

Secondary liability by way of contributory infringement arises in the United States when someone intentionally induces or encourages directly infringing acts, and arises by way of vicarious infringement by profiting from the directly infringing acts, while declining to exercise an ability to stop or limit it.[469] The contribution must induce, cause, or materially contribute to the infringing conduct of another,[470] which means that it must bear a direct relationship to the infringing acts.[471] The level of knowledge must be such that the defendant has reason to know that infringement is taking place.[472] In *Sony*, the U.S. Supreme Court found that suppliers of VHS recording devices, which could be used for recording television shows, were not liable for contributory copyright infringement because the devices were capable of substantial non-infringing use.[473] Specifically, the Court refused to impute the requisite level of knowledge from the characteristics or uses of the VHS recording devices as such, as they were capable of both infringing and substantial non-infringing use.[474] The *Sony* rule, which was inspired by the staple-article

243 (5th Cir. 2015); Rudkowski v. MIC Network, Inc., No. 17 CIV. 3647, 2018 WL 1801307, at *4 (S.D.N.Y. Mar. 23, 2018) ("An individual frame clearly represents an extremely small fragment of the whole Video.").

[466] For example, the European Court of Justice recently confirmed in *Pelham* that music sampling of very short phonograms would still count as an infringing reproduction, except where a sound sample from a phonogram was taken in order to use it, in a modified form unrecognizable to the ear, in a new work. *See* Case C-476/17, Pelham GmbH v. Ralf Hütter, ECLI:EU:C:2019:624, ¶¶ 29, 31 (July 29, 2019).

[467] *See supra* references cited in note 331.

[468] *See supra* references cited in note 331.

[469] Metro-Goldwyn-Mayer Studios Inc. v. Grokster, Ltd., 545 U.S. 913, 930 (2005).

[470] *Id.* at 930-31; Gershwin Pub. Corp. v. Columbia Artists Mgmt., Inc., 443 F.2d 1159, 1162 (2d Cir. 1971).

[471] Livnat v. Lavi, No. 96 CIV. 4967, 1998 WL 43221, at *3 (S.D.N.Y. Feb. 2, 1998).

[472] Sony Corp. of Am. v. Universal City Studios, Inc., 464 U.S. 417, 487 (1984).

[473] *Id.* at 442.

[474] *Id.*

doctrine from patent law,[475] has since been interpreted in *Napster* as meaning that the distribution of a commercial product capable of substantial non-infringing uses could not give rise to contributory liability for infringement unless the distributor had actual knowledge of specific instances of infringement and failed to act on that knowledge.[476] This means that more generalized knowledge of the possibility of a device being capable of being used for committing infringing acts is insufficient to attribute secondary liability. Instead, it is necessary that the contributory infringer has received notice of a specific infringing act from rightholders[477] or distributes a device with the object of promoting its use to infringe copyright as shown by clear expression or affirmative steps.[478] The original dataset provided by dataset creators will likely constitute a material contribution to the generative AI model in many cases. As discussed, the dataset is often fundamentally important to the model's function,[479] in which case dataset creators could become liable for contributory infringement. The pre-trained model that is completed which has also processed and pre-trained the original dataset is also obviously material to any infringing AI-generated work prompted by end-users.[480] The model is the necessary tool for creating AI-generated works and therefore directly facilitates any such infringements.

The more difficult question is whether dataset creators and/or model providers will possess the requisite knowledge that end-users of the generative AI model may be producing infringing works or that the model may be capable of doing so if given a particular prompt. Most generative AI-models are capable of substantial non-infringing uses, which means that the requisite knowledge cannot be derived as such from the characteristics of the model itself. It will be the exception rather than the norm that generative AI technologies are capable of producing "memorized" or near-verbatim copies of original works, and even when they do, it may be because the end user is intentionally trying to achieve this. As discussed previously in this Article, when researchers investigated how big of a problem "memorization" was

[475] *Id.* at 440-42.

[476] A & M Recs., Inc. v. Napster, Inc., 239 F.3d 1004, 1020-22 (9th Cir. 2001), *aff'd*, 284 F.3d 1091 (9th Cir. 2002) (holding that Napster was not secondarily liable for copyright infringement committed by its users until it was on notice of specific infringing acts that it failed to prevent). In *Aimster*, the Seventh Circuit also found that there could be contributory infringement where a file-sharing company had "willful blindness" of primary infringements committed by users of their platform, which the Court equated with "knowledge" of infringing activity. *See In re* Aimster Copyright Litig., 334 F.3d 643, 650 (7th Cir. 2003).

[477] A & M Recs. Inc., 239 F.3d at 1020-22. However, case law is not conclusive, and some other courts have come to adopt their own, lesser variations of a knowledge requirement, including a should-have-known standard. *See* Religious Tech. Ctr. v. Netcom On-Line Comm., 907 F. Supp. 1361, 1374 (N.D. Cal. 1995) (holding that the question was whether the defendant should have known whether there was primary infringement); Arista Records, Inc. v. Flea World, Inc., No. CIV.A. 03-2670, 2006 WL 842883 at *15 (D.N.J. Mar. 31, 2006) (holding that the contributory infringer, a flea market operator, knew or should have known of primary infringement ongoing in the market).

[478] Metro-Goldwyn-Mayer Studios Inc. v. Grokster, Ltd., 545 U.S. 913, 936-37 (2005).

[479] *See supra* Section II.

[480] *See* Alhadeff et al., *supra* note 274, at 35; Lee et al., *supra* note 7, at 97.

for Stable Diffusion, they found that only about 1.2% of 100,000 randomly sampled user-generated captions were potentially sufficiently similar to the original training data, which would indicate that they may be duplicates.[481] In another survey, when researchers randomly sampled LLMs from the GPT-Neo model family, which is an open-source LLM released by EleutherAI, they found that extractable memorization was at least 1% of the training dataset.[482] However, this does not mean that the memorization rate can be significantly higher in certain more specialized use cases or when additional data is used for fine-tuning.[483] What will also be lacking in most cases is the actual knowledge of the infringing activity. Neither dataset creators nor model providers will typically have actual knowledge of limited instances of any infringing activity until they receive notice from rightholders.[484] This altogether suggests that contributory infringement may be a rather precarious argument against dataset creators and model providers, at least based on the law as it currently stands.

The secondary liability position is even more difficult in other countries. For example, in the U.K., there is no general doctrine of contributory copyright infringement. Instead, there are a limited number of narrowly defined circumstances set out in statute where someone can become indirectly liable for someone else's primary copyright infringement. One such statutory provision is Section 22 of the CDPA 1988, which provides that "copyright in a work is infringed by a person who, without the license of the copyright owner, imports into the United Kingdom, otherwise than for his private and domestic use, an article which is, and which he knows or has reason to believe is, an infringing copy of the work." If the

[481] Somepalli et al., *supra* note 253, at 2. Results from other studies are not fully conclusive. In another study, Carlini et al. found that only 107 near-duplicate images of training data were found when sampling Stable Diffusion v1.4, based on a selection of 350,000 of the most duplicated examples from the training dataset (which were used to generate a total of 175 million images). *See supra* note 57, at 4-6. This would equate to a memorization rate of approximately 0.000061%, in other words significantly lower than other studies. However, Carlini et al. used an extraction methodology to more accurately identify false positives, as opposed to false negatives, *see id.*, which might explain the very low results.

[482] Nicholas Carlini et al., *Quantifying Memorization Across Neural Language Models*, ARXIV 1 (Mar. 6, 2023), https://arxiv.org/pdf/2202.07646 [https://perma.cc/82SY-BDK7]; *See also id.* at 7 (finding an extractable memorization rate of just under 2% on average when randomly sampling 100,000 sequences).

[483] For example, when various open-source LLMs were used, a memorization rate was found to be as high as 20% for fine-tuned data from medical dialogs. *See* Shenglai Zeng et al., *Exploring Memorization in Fine-tuned Language Models*, ARXIV 3 (Feb. 22, 2024), https://arxiv.org/pdf/2310.06714 [https://perma.cc/W2P8-Z8KW] ("For summarization and medical dialog models, we identified total memorization rate of 20.7% and 19.6%, respectively.").

[484] Lee et al., *supra* note 7, at 97. As discussed, Article 53(d) of the EU AI Act, *supra* note 11, now provides that providers of general-purpose AI models, but not any other model or AI system, has an obligation to draw up a sufficiently detailed summary of the training data used. *See supra* Section III.B.2.v. This will provide rightholders with the possibility, at least in theory, to vet training datasets for their works and provide adequate notice to dataset creators and/or model providers, however it is unclear at this point how this obligation will be implemented in practice. A similar transparency obligation will arise for generative AI systems deployed in California under the recently passed bill. *See* A.B. 2013, 2023-2024 Leg., Reg. Sess. (Cal. 2024).

training data is encoded into the model parameters, and if the model is subsequently marketed and sold in the country at issue, then there is an argument that the model contains infringing "works" which have been "imported."[485] This was too raised by Getty Images in its pending infringement complaint against Stability AI,[486] with issues surrounding whether intangible information can be considered an "article" that can be imported. However, and as discussed above,[487] the argument also fails if the training data is not actually encoded into the model parameters, such that it can represent an "infringing copy" which can be "imported." This is a complex technical question that will depend on the facts and would require an expansive statutory reading. Generally, there are significant difficulties with finding that individual and specific works have been reproduced into the model itself, thereby also turning the model into an "infringing copy."

In the EU, secondary liability in case of copyright infringement is not harmonized and approaches differ significantly at a national level, which falls outside the scope of this Article to cover.[488] What is harmonized, however, at an EU-level, is the safe harbor provision in the e-Commerce Directive[489] for online intermediaries. Specifically, Article 14 of the Directive provides that, where information society service is provided that consists of the storage of information provided by a recipient of the service, the provider will not be liable for the information stored at the request of a recipient of the service if they have no "actual knowledge of illegal activity or information" and "[are] not aware of facts or circumstances from which the illegal activity or information is apparent."[490] Alternatively, such providers are not liable if upon obtaining such knowledge or awareness, they "act[] expeditiously to remove or to disable access to the information."[491] This provision is more relevant for online marketplaces or file-hosting or file-sharing platforms as it presumes there is an "intermediary" service provider.[492] This is clearly not the case in the present context of generative AI

[485] *See supra* Section III.C.2 (discussing how encoding copyrighted materials into model parameters, or "memorization," may amount to a separate act of reproduction).

[486] Getty Images (US) Inc v. Stability AI Ltd [2023] EWHC (Ch) 3090 (U.K.).

[487] *See supra* Section III.C.2.

[488] *See* JAN BERND NORDEMANN, DIRECTORATE GENERAL FOR INTERNAL POLICIES, EUROPEAN PARLIAMENT, LIABILITY OF ONLINE SERVICE PROVIDERS FOR COPYRIGHTED CONTENT – REGULATORY ACTION NEEDED? 5 (2021), https://www.europarl.europa.eu/RegData/etudes/IDAN/2017/614207/IPOL_IDA(2017)614207_EN.pdf [https://perma.cc/3N39-7QEG]. For a more comprehensive survey about these differences between Member States, see Christina Angelopoulos, *Harmonising Intermediary Copyright Liability in the EU: A Summary*, *in* THE OXFORD HANDBOOK OF ONLINE INTERMEDIARY LIABILITY 315 (Giancarlo Frosio ed., 2020).

[489] Council Directive 2000/31, 2000 O.J. (L 178) 1 (EU) [hereinafter e-Commerce Directive].

[490] *Id.* at art. 14(1)(a).

[491] *Id.* at art. 14(1)(b). This has been interpreted by the European Court of Justice as meaning that awareness of facts or circumstances on the basis of which a diligent economic operator should have identified the illegality is sufficient to meet the knowledge standard. *See* Case C-324/09, L'Oréal SA v. eBay, 2011 E.C.R. I-6073, ¶¶120-24.

[492] *L'Oréal SA*, 2011 E.C.R., ¶¶ 111-13; Joined Cases 682 & 683/18, Frank Peterson v. Google LLC, ECLI:EU:C:2021:503, ¶¶104-106 (June 22, 2021).

models,[493] where datasets are created without the participation of end-users. Nor are dataset creators or model providers merely providing a "mere conduit" service or transmission[494] or caching.[495] The e-Commerce Directive would therefore need to be amended if the EU wants to extend it to dataset creators or model providers. In the absence of any such safe harbor provision, dataset creators or model providers could potentially become secondarily liable for copyright infringement under national copyright law or tort-based rules in different Member States.

However, even if we would consider that dataset creators or model providers should be secondarily liable for infringing AI-generated works produced by end-users of the model, a significant limitation with any rule like this is that secondary liability is reserved for cases of primary infringement. As discussed, it is expected that primary infringements committed by end-users of AI models will be the exception rather than the norm in most cases.[496] If the argument above fails and AI-generated works are not infringing derivative works, then so will the argument fail that dataset creators or model providers contribute to any infringement. Without a primary infringement, there can be no contributory infringement. Either policymakers will consider this limited liability to be sufficient to protect rightholders or they will not. If they will not, then policymakers have to go back to square one and critically consider how the inputted training data, comprising copyrighted materials, can be regulated in a way that closes the transnational data loophole while respecting foreign sovereignty. This Article has concluded that the best way to do that is to revise existing international copyright treaties to permit states to extraterritorially regulate wholly foreign training activities. Alternatively, states should come to agreement on internationally harmonized exceptions and limitations for text and data mining.

IX. PRACTICAL CONSIDERATIONS FOR AI DEVELOPERS AND PROVIDERS

As AI regulations are rapidly evolving around the world and as courts start interpreting the extent of copyright laws on data reproduction in the course of model training, AI model developers are put in a difficult position. It is likely that the same act of reproduction will be infringing on copyright in some countries, and not infringing in others, depending on the individual circumstances. There are presently some countries, such as the U.K.,[497] Australia,[498] and India,[499] where there are no exceptions or limitations in place that could readily prevent copyright infringement. A common-sense approach would be to move the training activities and the

[493] *L'Oréal SA,* 2011 E.C.R. I-6073, ¶113.

[494] Council Directive 2000/31, art. 12, 2000 O.J. (L 178) 1 (EU)

[495] *Id.* at art. 13.

[496] *See supra* Section VIII.A (discussing how AI-generated works will only exceptionally be so similar to the original training data such that they will primarily infringe, for example when the training data has been "memorized" into the model parameters).

[497] *See supra* Section III.B.3.

[498] *See supra* Section III.B.4.

[499] *Id.*

business operations to a country with the greatest legal certainty and more AI-friendly copyright rules, as OpenAI itself has foreshadowed that it might do. But even that may not suffice. At least the EU position now appears to be that it will extraterritorially apply the opt-out procedure set out in the Copyright Directive to foreign training activities *if* the model is later placed on the European market.[500] The only good news for developers is that the EU regime is relatively predictable compared to its contemporaries. There are several points that lack clarity when it comes to interpreting and implementing the opt-out procedure in the Copyright Directive,[501] but notwithstanding that the rule is straightforward in the sense that, if a valid opt-out notice has been provided, then the material must be excluded from the training dataset. There is no case-by-case analysis, which is the case for the fair use exception in the United States and which is highly impractical for developers.

It could spell trouble if other jurisdictions follow the same footsteps as the recitals in the EU AI Act and extend their own copyright laws for AI development extraterritorially. Unlike other regulatory domains, like data privacy, where compliance with extraterritorial provisions can be managed through more straightforward legal and technical adjustments, AI presents unique challenges. A model is typically only developed once and requires substantial investment, particularly LLMs where development costs may exceed many billions of dollars.[502] Those costs extend to computational resources,[503] electricity costs,[504] data acquisition,[505] and human labor involved in curating and refining training datasets. The potential for regulatory fragmentation and associated costs are substantial, which should be recognized. One of the promises of generative AI is that it is currently cheap compared to traditional creative services. But if development costs escalate many times more than what is already the case, that price tag may well materially change. If regulators or courts follow the same

[500] *See supra* Section VII.A.1 (discussing how the EU AI Act, *supra* note 11, extraterritorially extends the opt-out procedure in Article 4(3) of the Copyright Directive, *supra* note 117).

[501] *See supra* Section III.B.2 (discussing that it is unclear, for example, when rightholders are deemed to have duly "opted-out," and how dataset creators and managers are supposed to deal with inadvertently infringing materials, derived from third parties and which form part of vast datasets).

[502] Craig S. Smith, *What Large Models Cost You – There Is No Free AI Lunch*, FORBES, (Sept. 8, 2023), https://www.forbes.com/sites/craigsmith/2023/09/08/what-large-models-cost-you--there-is-no-free-ai-lunch/ [https://perma.cc/EKK6-VVRS] (quoting Sam Mugel, Chief Technology Officer of Multiverse Computing, who estimates that "training the next generation of large language models will pass $1 billion within a few years"). Meta, for example, recently also announced that it plans to spend between $37 billion and $40 billion in 2024 on AI infrastructure. *See* Ana Altchek, *Meta's Going to Spend Like Crazy on AI in the Next Year — And Investors Don't Hate It*, BUS. INSIDER (July 31, 2024), https://www.businessinsider.com/meta-ai-spending-more-mark-zuckerberg-investors-stock-price-2024-7 [https://perma.cc/LZR6-3W98].

[503] *See* Smith, *supra* note 502 ("The cost of the GPUs, alone, can amount to millions of dollars. According to a technical overview of OpenAI's GPT-3 language model, each training run required at least $5 million worth of GPUs").

[504] *See* Siddharth Samsi et al., *From Words to Watts: Benchmarking the Energy Costs of Large Language Model Inference*, ARXIV (Oct. 4, 2023), https://arxiv.org/pdf/2310.03003 [https://perma.cc/5T9P-YXZ6]

[505] This includes procuring license agreements for accessing copyrighted works, publicly sourced datasets, as well as synthetic data from data providers and marketplaces.

approach by applying their copyright laws extraterritorially, then there is also a risk that developers are forced to choose between abandoning some domestic markets entirely or releasing less potent and less powerful AI models using exclusively public domain and/or authorized datasets. As discussed elsewhere in this Article,[506] some developers have already prepared for the latter. For example, OpenAI has recently struck licensing deals with content producers and aggregators like Associated Press, Axel Springer, Financial Times, Reddit, Vox Media and Shutterstock, among others.[507]

For now, most developers are likely better off adopting a more cautious "wait-and-see" approach, in particular considering that there would be substantial development costs associated with creating models tailored to each country's copyright laws. It is presently not clear that an extraterritorial application of copyright laws, to which acts of reproduction are exclusively foreign, is permitted.[508] There is a possibility that European courts will decline to give legal force to the far-reaching statements that can only be found in the EU AI Act recitals, especially where there is a substantial risk of conflicting with foreign copyright laws and interests.[509] There is also the possibility that no other jurisdiction will take the same approach. But if the same trend that followed the data privacy laws after the GDPR is to be indicative,[510] then there is a good chance that other countries will come to adopt similar extraterritoriality provisions in their copyright laws, which Brazil has so far proposed doing, either to protect their rightholders' interests or to preserve, as the EU put it, "a level playing field among providers of general-purpose AI models."[511]

Until then, it makes commercial sense for developers to offshore their training activities to AI-friendly jurisdictions where this is clearly not infringing on copyright. At the time of writing this Article, this may be Singapore, which appears to have the most permissible text and data mining exception.[512] The second most permissible jurisdiction would be the EU. However, because the text and data mining exception in the EU may be applied extraterritorially to foreign training activities, this means in practice that EU becomes the preferred location worldwide

[506] *See supra* Section III.B.1.

[507] *AI Content Licensing Deals: Where OpenAI, Microsoft, Google, and Others See Opportunity*, CBINSIGHTS (July 19, 2024), https://www.cbinsights.com/research/ai-content-licensing-deals/ [https://perma.cc/RS3R-N8SD].

[508] *See supra* Section VII.A.4 (discussing how extraterritorially extended exceptions and limitations for text and data mining might unreasonably interfere with foreign sovereignty and run counter to Article 5(2) of the Berne Convention, *supra* note 92).

[509] *Id.*

[510] *See supra* references cited in note 13. It should be noted that Singapore that adopted extraterritoriality provisions in their data privacy laws have so far not done the same with respect to copyright. However, this might be explained by the fact that there is no opt-out procedure in the Singaporean text and data mining exception, and because of that, there is no significant obligation that could be avoided by relocating training activities to another country. *See supra* Section III.B.4.ii (Singapore).

[511] EU AI Act, *supra* note 11, at Recital 106.

[512] *See supra* Section III.B.4.ii (Singapore).

to situate training activities if there is a commercial interest in any event to release the model on the EU market. For example, if training activities are relocated to Singapore, and if the completed model is later placed on the EU market, then *both* the EU and Singaporean text and data mining copyright rules will come to apply. In contrast, if the training activities would instead be relocated to the EU, then *only* EU copyright law would apply. That is, of course, assuming other countries do not follow the same trend of copyright extraterritoriality as the EU.[513] Many companies would prefer to only follow one set of laws that target something so fundamental as model data training than multiple potentially divergent laws. The extraterritorial reach of EU copyright law could therefore mean that it will not apply to many developers if they decide to relocate their training activities in any event to the EU. Ironically, extraterritoriality may drive developers right back to the EU's doorstep.

The copyright situation is generally similar for service providers of AI tools or for custom developers engaged in fine-tuning on more specific datasets. There is a risk that these actors will also be considered to infringe on copyright if they are using open-source LLMs that have been trained on infringing datasets if the original data is being encoded into the default model parameters.[514] As discussed, the exclusive right of reproduction in copyright law extends to acts of copying of original works "in any manner or form."[515] It is a complex technical question whether original training data becomes imprinted into the model itself, which is expected to evolve and become more clear as lawsuits turn into decisions. If encoded and memorized original content also infringes copyrighted works, then custom developers using open-source LLMs must prepare for the same complexities above. In addition, custom developers must carefully vet the more specific datasets they employ themselves to ensure compliance with applicable copyright laws. Custom developers using RAG or similar technologies potentially face greater copyright constraints in this regard, as these rely on extracting original content from sources. If the extracted content does not qualify as a permitted quotation under applicable copyright law, and if no other exceptions or limitations apply, then these practices could be directly infringing.

No matter what jurisdiction developers or providers are in, it is clear that the training dataset must not contain infringing works.[516] What is not clear is what developers actually need to do to prevent that from happening, and there is a high likelihood that courts in different jurisdictions will take different approaches in that regard. The issue that is expected to be answered differently between courts is how to tackle the situation where developers have no easy means of knowing or confirming that infringing content has been included in the massive original training dataset, which may sometimes come from third party sources. This Article

[513] *See supra* Section VII.A.2 (discussing the international trends witnessed so far regarding copyright extraterritoriality for text and data mining).

[514] *See supra* Section III.C*(*discussing to what extent training data encoding into model parameters could infringe on copyright*)*.

[515] Berne Convention, *supra* note 92, art. 9(1). *See supra* Section III.A.

[516] *See supra* Section III.B.2 (focusing the discussion on the specific rules that apply in the EU in this regard).

has proposed, as a means of solving that dilemma, to introduce a knowledge requirement similar to what EU courts have adopted in the case of hyperlinking.[517] In practice, this would mean that dataset creators or model developers should only become liable for infringing content contained in the dataset if they possess actual, imputed, or constructive knowledge of infringement. Until this becomes clarified in case law, dataset creators and developers are wise to invest in developing vetting routines or technologies that can detect potentially infringing works from vast quantities of data. Another pragmatic, and more affordable, solution would be to exclude data from sources known to more frequently contain infringing materials and willingly assume the legal risk that other sources could potentially be compromised in rarer cases. Alternatively, they can invest in procuring licenses from authorized dataset providers with appropriate warranties and indemnities or obtain legal insurance from insurance providers.

A very important limitation of the extraterritorial extension of the opt-out procedure for text and data mining in the Copyright Directive via the EU AI Act is that it only applies, according to its wording, to providers of general-purpose AI models.[518] This Article has discussed that this potentially creates a loophole in the rules, if general-purpose AI models are never provided directly to the EU market but only indirectly via general-purpose AI systems which are based on such models.[519] It is presently unclear if providers of general-purpose AI systems from third countries have to comply with the same extraterritoriality rules.[520] What makes even less sense is that providers of AI systems, which are separate from general-purpose AI models, who are established in third countries do not have to comply with the text and data mining exception in the Copyright Directive. This creates an unequal playing field between providers of AI systems who are established in the EU and those who are not. The former have to comply directly with the opt-out procedure in text and data mining exception in Article 4(3) of the Copyright Directive, but the latter will have no such obligation. This distinction is not commented on, let alone explained, in the EU AI Act. In practice, this critically means that developers of AI systems could avoid the obligation to respect rightholders' opt-outs from text and data mining if they relocate their training activities to outside the EU.

[517] *See supra* Section III.B.2(referring to Case C-160/15, GS Media BV v. Sanoma Media Neth. BV, ECLI:EU:C:2016:644, ¶49 (Sept. 8, 2016)). This would also be similar to Copyright Act 2021, No. 22 of 2021, §244(2)(e)(ii)(2021) (Sing.), and the safe harbor rule for online intermediaries that provide storage services in the e-Commerce Directive, *supra* note 489, at art. 14 (EU). *See supra* Section VIII.B.

[518] EU AI Act, *supra* note 11, at Recital 106, art. 53(1)(c).

[519] *See supra* Section VII.A.1 (explaining that Recital 106 and Article 53(1)(c) of the EU AI Act, *supra* note 11, by their wording, only applies to providers of general-purpose AI models, but not general-purpose AI systems).

[520] *See supra* Section VII.A.1.

X. LEGAL CONSIDERATIONS FOR RIGHTHOLDERS

Rightholders presently have many more options in their litigation arsenal than AI developers and providers. The first and obvious choice is to sue for copyright infringement in the country where the servers onto which training data is reproduced and stored are situated. If that is unfavorable, either because it is unknown where the servers are located or because the servers are located in an AI-friendly jurisdiction, then the second choice should be to sue for infringement where the developers or providers are located. This poses the risk, however and as discussed above,[521] that courts will deny the claim out of extraterritoriality concerns. Because the act of reproduction literally takes place where the data is reproduced, and if the data is stored on servers located somewhere else, some courts may be reluctant to extend their copyright laws to such conduct. Suing for copyright infringement at the principal place of business should therefore only be a last litigation resort.

The situation becomes more complex where AI developers and providers have moved their entire operation, both servers and business, to more AI-friendly countries. If the allegedly infringing conduct revolves around reproduction during training, and if the act of reproduction takes place in its entirety abroad, courts will be rightfully reluctant to extraterritorially extend the reach of domestic copyright laws.[522] Even if the model, once completed, is later marketed and sold to domestic customers, the issue is that the relevant infringing conduct is focusing on the underlying development process, which takes place somewhere else, and not the resulting product. New legal theories would need to be submitted to close that loophole. Recital 106 and Article 53.1(c) of the EU AI Act is one such example where copyright exceptions and limitations for text and data mining are applied extraterritorially in this regard. Yet these rules are unclear.[523] It is possible that they should be interpreted not to extend copyright law as such extraterritorially, in which case the only remedy for not complying with the rules would be non-compliance sanctions under the EU AI Act.[524] Those sanctions will benefit public authorities, while rightholders will ironically not receive a penny.

[521] *See supra* Section V (discussing how it is not clear whether courts will localize foreign acts of reproduction to the place where the party instructing such acts resides).

[522] *See supra* Section VI (explaining that, if no act of reproduction, in whole or in part, is deemed to occur within the relevant country of protection, then there can be no copyright infringement traceable to the model training activity).

[523] *See supra* Section VII.A.1 (discussing how, for example, the EU AI Act, *supra* note 11, does not make it clear whether EU copyright law is extraterritorially extended as such, or whether only its provisions are extraterritorially extended in a way which mirrors EU copyright standards, specifically Article 4(3) of the Copyright Directive, *supra* note 117). Recital 106 and Article 53(1)(c) of the EU AI Act, *supra* note 11, also only apply to providers of specifically general-purpose AI models, meaning that providers and deployers of AI systems will not have to comply extraterritorially with opt-outs from rightholders if conducting their training activities outside the EU.

[524] *See supra* Section VII.A.1.

However, to the extent that the EU AI Act should be interpreted to, in fact, extend copyright law extraterritorially, there is also a high likelihood that this could interfere with foreign sovereignty and be incompatible with Article 5(2) of the Berne Convention.[525] Because of those difficulties, this Article has discussed the possibility of comparing how patent law deals with products derived from patented processes or methods.[526] It is well-established in international patent law that courts may extend the reach of domestic patent laws to prevent the importation of products derived from patented processes, which have been undertaken abroad, even if there is no patent protection in that foreign country.[527] Similar arguments could be raised in the copyright and AI context, which is similarly a product-by-process problem, although it would first require that copyright laws are amended in the same way as patent laws. This type of legislative intervention in copyright law would make a lot of sense to rightholders. As discussed throughout this Article, what rightholders are mostly concerned about when it comes to generative AI is not necessarily the fact that they lose out on licensing fees for the reproduction of their works in model training, which may be comparatively small in total numbers. What rightholders so far have raised their voices about the most is how generative AI works could threaten and directly compete with the original work that they produce themselves.[528] Applying a product-by-process legal framework to model training is attractive for that reason. It accounts for the fact that, just like patent law did for method or process patents, it is the generated output, not the input training data, which is most economically valuable.

Another enforcement option comes into play if the training data is encoded onto the model parameters themselves. If that is the case and can be proven, then rightholders may decide to go after whoever is reproducing the model parameters, wherever that may occur. That could be not just developers or model providers, but also custom developers that fine-tune the models or other actors who require access to the model parameters for specific deployment. One argument currently pending review in *Getty Images v. Stability AI* in the U.K., is whether secondary infringement provisions can provide recourse in this regard.[529] This revolves around the question of whether the model is an "article" containing "infringing

[525] *See supra* Section VII.A.4 (discussing, in particular, how extraterritorially extended exceptions and limitations for text and data mining could mean that copyright is no longer "exclusively" governed by domestic copyright law, as required by Article 5(2) of the Berne Convention, *supra* note 92).

[526] *See supra* Section VII.B (discussing how AI-generated works derived from generative AI models could be compared with products derived from patented processes in patent law).

[527] *See supra* Section VII.B (referring to Article 5 *quater* of the Paris Convention, *supra* note 435, which permits the extraterritorial application of national patent law to imported products derived from foreign processes, to the extent that a particular country has introduced such protection in their own patent statute).

[528] *See* complaints cited *supra* note 440.

[529] *See supra* Sections III.C.2, VIII.B.

copies" of original works, which is a difficult technical analysis for reasons stated earlier.[530]

The last available recourse is to target the output content for copyright infringement. Generated output may infringe either directly, where the output is substantially similar or identical to original content or indirectly, where the output is derivative of original content.[531] If either output is infringing, then dataset creators and model providers might be secondarily liable for contributory infringement for the primary acts taken by users.[532] This is, however, necessarily a case-by-case analysis, which will largely depend on the facts, as well as which jurisdiction is concerned. Generally speaking, there may be relatively few AI-generated works that will primarily infringe on the copyright of original training data. Although there is a need for additional technical studies that look into the issue, studies conducted so far suggest that it will be the exception rather than the norm that generative AI models "memorize" the original training data, such that the model is capable of producing an identical or near-identical version of that data.[533] However, that does not mean that there could be some training datasets that have a greater proportion of "memorized" training data. Some studies suggest that this may be the case for datasets used for fine-tuning[534] or where certain training data becomes duplicated a significant number of times in the dataset.[535] To the extent there is "memorized" infringing data in the model parameters, then dataset creators and model providers will directly infringe whenever they reproduce the model parameters. But it also adds the risk for dataset creators and model providers to become secondarily liable for the primary copyright infringement committed by end-users upon prompting that "memorized" infringing content.[536] It is presently unclear to what extent dataset creators and model providers could become secondarily liable in these circumstances, and at least in the United States, it may be a relatively precarious argument.[537] Failing that argument, rightholders will need to go back to square one and consider the enforcement options for targeting the

[530] *See supra* Section III.C.2. *See also supra* Section III.C.1 (discussing how technically difficult it is to prove what is going on inside generative AI models, and to what extent training data is becoming "memorized" in model parameters).

[531] *See supra* Section VIII.A (discussing whether generated output could count as infringing derivative works).

[532] *See supra* Section VIII.B (discussing to what extent dataset creators and model providers could become secondarily liable for primary infringements from end users).

[533] *See supra* Section VIII.B. *See also supra* Section III.C.2 (with further references on how significant a problem "memorization" of original training data is).

[534] *See supra* Section VIII.B (referring to Zeng et al., *supra* note 483, at 3, discussing a high memorization rate of up to 20% for specialized datasets used for fine-tuning).

[535] Somepalli et al., *supra* note 253, at 9 (discussing how not only image but also caption duplication makes it more likely that an image will become memorized in the dataset).

[536] *See supra* Section VIII.B.

[537] *See supra* Section VIII.B (discussing that, because dataset creators and model providers will often not have actual knowledge of specific infringing acts, and because generative AI models are capable of substantial non-infringing use, the risk of contributory copyright infringement is relatively limited, as the law currently stands).

underlying training activity that takes place, assuming this occurs in a country where it amounts to copyright infringement.

XI. THE FUTURE OF AI GOVERNANCE AND DIGITAL SOVEREIGNTY

The EU AI Act is the first comprehensive AI regulatory framework in the world. It is also the first time in history that copyright law, whether directly or indirectly, has come to be extraterritorially extended to what is exclusively foreign conduct, specifically text and data mining activities for general-purpose AI models.[538] It is not clear whether other countries will follow the same footsteps as the EU in this regard. So far, only Brazil has proposed similar rules that extraterritorially extend Brazilian copyright law to foreign training activities.[539] If this trend of copyright extraterritorially continues for model training, then AI developers and providers will face a very serious problem. AI developers and providers will be forced to comply with multiple diverging copyright laws if they end up releasing their model in the countries which have enacted such extraterritoriality provisions that will come to apply retroactively to the same underlying process. This could undermine the entire AI industry.

In the worst-case scenario, as discussed above,[540] the extraterritorial application of copyright laws against foreign training activities could result in fewer internationally available AI models. AI developers may decide to focus on larger markets with clear and established copyright exceptions and limitations and abandon other markets with extraterritoriality provisions due to increased development costs from stricter copyright constraints. In the best-case scenario, we may come to see different variations of AI models which are available in different markets.[541] This assumes, however, that AI developers have the financial and technical muscle, and sufficient data, to rework the AI development lifecycle on a country-by-country basis. AI development, particularly larger models, is an enormously expensive and complex process and represents a very significant investment.[542]

But copyright law is just a piece of a much larger puzzle. AI regulations will impose a patchwork of obligations on both developers and providers. These rules will inevitably come to differ in the details between jurisdictions without international harmonization. The obligations on AI developers and providers in the EU AI Act are comprehensive, particularly for AI systems classified as high-risk. High-risk AI systems under the EU AI Act must establish a risk management

[538] EU AI Act, *supra* note 11, at Recital 106, art. 53(1)(c).

[539] *See supra* Section VII.A.2 (discussing Brazil's proposed AI regulations, which include provisions on copyright text and data mining exceptions with certain requirements to satisfy, insofar as research organizations and institutions are concerned, and which include provisions that require other AI developers to remunerate rightholders for their text and data mining activities).

[540] *See supra* Section VII.A.3 (discussing how copyright extraterritoriality could undermine the AI industry through excessive concurrent jurisdiction).

[541] *See supra* Section VII.A.3.

[542] *See supra* Section IX (discussing, with further references, the significant costs associated with AI development).

system, draw up detailed technical documentation to demonstrate compliance, implement human oversight measures, and establish a quality management system among other obligations.[543] Article 10 of the Act also sets out rules for data governance, including what testing datasets may be used for high-risk AI systems. In case ChatGPT would be classified as high-risk under the EU rules, OpenAI's CEO Sam Altman has said that "[e]ither we'll be able to solve those requirements or not . . . . If we can comply, we will, and if we can't, we'll cease operating… We will try. But there are technical limits to what's possible."[544] The EU AI Act also introduces more stringent rules for providers of general-purpose AI models with systemic risk.[545] These include obligations to perform model evaluation, systemic risk mitigation and ensuring an adequate level of cybersecurity protection.[546] Providers of general-purpose AI models that are not systemic risk the need to comply with fewer obligations, such as drawing up technical documentation and making available drawing up a summary of the training data used.[547]

In addition to burdening AI regulations and copyright concerns, data privacy laws also impact transnational AI development and deployment. Like copyright, data privacy restrictions are substantial because they touch the underlying training data. The processing of personal data typically requires a legal basis under data privacy laws such as the GDPR.[548] Where it is impractical, if not impossible, to retroactively obtain consent at scale from unknown individuals whose personal data may be part of the training dataset, developers will need to rely on other legal basis such as legitimate interests.[549] While measures like pseudonymization can sometimes be used to remove personal data from the source material, it is not foolproof, and data may still be re-identified when combined with other datasets containing identifiable information. Data privacy hit the headlines back in April 2023, when Italy temporarily banned ChatGPT over data privacy concerns, stating that there was no legal basis to justify "the mass collection and storage of personal data for the purpose of 'training' the algorithms underlying the operation of the platform."[550] The ban was eventually withdrawn a month later after OpenAI made

[543] EU AI Act, *supra* note 11, at arts. 8-17.

[544] Billy Perrigo, *OpenAI Could Quit Europe Over New AI Rules, CEO Sam Altman Warns*, TIME (May 25, 2023), https://time.com/6282325/sam-altman-openai-eu/ [https://perma.cc/U5YC-N265].

[545] Article 3(65) of the EU AI Act, *supra* note 11, defines "systemic risk" as a risk that is specific to the high-impact capabilities of general-purpose AI models, having a significant impact on the Union market due to their reach, or due to actual or reasonably foreseeable negative effects on public health, safety, public security, fundamental rights, or the society as a whole, that can be propagated at scale across the value chain.

[546] *Id.* at art. 55(1).

[547] *Id.* at art. 53(1).

[548] GDPR, *supra* note 13, art. 6(1) (EU).

[549] *Id.* at art. 6(1)(f). *See also* Pablo Trigo Kramcsák, *Can legitimate interest be an appropriate lawful basis for processing Artificial Intelligence training datasets?*, 48 COMPUT. L. & SEC. REV. 1 (2023) (concluding that legitimate interests may better match, than obtaining individual consent, for the purpose of building AI training datasets).

[550] Shiona McCallum, *ChatGPT banned in Italy over privacy concerns*, BBC (Apr. 1, 2023) https://www.bbc.com/news/technology-65139406 [https://perma.cc/4NLE-7D84].

assurances that everyone had a right to opt-out of the processing of their personal data by an easily accessible online form.[551] To make matters worse, and as has been mentioned above,[552] data privacy laws frequently have extraterritorial provisions, forcing AI developers and providers to navigate across conflicting requirements.

The enforcement of copyright, AI and data privacy regulations extraterritorially is done in the name of digital sovereignty. States are sovereign over what is happening within their borders, not just physically but also digitally.[553] Digital and online platforms, including artificial intelligence, are being developed in other countries yet are capable of being accessed domestically, impacting the digital and societal landscape. The European Parliament has responded by stating that "[digital service platforms and artificial intelligence] have redefined how we communicate, shop and access information online, making them daily essentials. The European digital agenda for 2020-2030 addresses these shifts. It prioritizes establishing secure digital spaces, ensuring fair competition in digital markets and enhancing Europe's digital sovereignty ."[554] Indeed, Ursula von der Leyen stated in a speech before the European Commission back in 2019 that "[w]e must have mastery and ownership of key technologies in Europe. These include quantum computing, *artificial intelligence*, blockchain, and critical chip technologies."[555] The EU AI Act is one such example of the EU trying to maintain its sovereignty over AI as a critical technology.[556]

Many more countries are expected to follow in the same or similar footsteps as the EU when it comes to regulating AI, and there are now more than 1000 policy

[551] *Italy lifts ban on ChatGPT after data privacy improvements*, DW (Apr. 29, 2023), https://www.dw.com/en/ai-italy-lifts-ban-on-chatgpt-after-data-privacy-improvements/a-5469742 [https://perma.cc/QWG7-CNLG].

[552] *See supra* note 13 (quoting, in addition to the GDPR, the data privacy laws of California, Singapore, Australia, Brazil, Canada, and China).

[553] *See, e.g.*, Paul Timmers, *Sovereignty in the Digital Age*, *in* INTRODUCTION TO DIGITAL HUMANISM 571 (Hannes Werthner et al. eds., 2023); Julia Pohle & Thorsten Thiel, *Digital Sovereignty, in* PRACTICING SOVEREIGNTY: DIGITAL INVOLVEMENT IN TIMES OF CRISES 47-61 (Bianca Irrgang Herlo et al. eds., 2021); Huw Roberts, Emmie Hine & Luciano Floridi, *Digital Sovereignty, Digital Expansionism, and the Prospects for Global AI Governance*, *in* QUO VADIS, SOVEREIGNTY? 51 (Phil. Stud. Series No. 154, 2023); Jack L. Goldsmith, *The Internet and the Abiding Significance of Territorial Sovereignty*, 5 IND. J. GLOB. LEGAL STUD. 475-91 (1998).

[554] *Digital Agenda for Europe*, EUROPEAN PARLIAMENT: FACT SHEETS ON THE EUROPEAN UNION (May 2024), https://www.europarl.europa.eu/factsheets/en/sheet/64/digital-agenda-for-europe [https://perma.cc/E83M-82XT]. *See also* Andrea Calderaroa & Stella Blumfelde, *Artificial intelligence and EU security: the false promise of digital sovereignty*, 31 EUROPEAN SEC. 415, 417-20 (2022) (summarizing the European approach to digital sovereignty).

[555] Ursula von der Leyen, President-elect, European Commission, Speech on the Occasion of the presentation of her College of Commissioners and their programme (Nov. 27, 2019), https://ec.europa.eu/commission/presscorner/detail/es/speech_19_6408 [https://perma.cc/S6SH-TV29] (emphasis added).

[556] Other recent regulatory initiatives from the EU include the EU Data Act, Regulation 2023/2854, 2023 O.J. (L 71) 1, which sets out rules on fair access and use of data between businesses, users and the public sector, particularly within the Internet-of-Things framework.

initiatives pending in more than seventy countries, and counting.[557] For example, Brazil, which has been discussed already,[558] and Canada[559] have announced similar draft AI regulations that introduce more onerous requirements for AI systems based on certain risk thresholds. The United States has no federal AI regulations as of yet, and with no federal initiatives pending, individual states are taking the lead.[560] Colorado became the first state in May 2024 to adopt specific AI regulations,[561] which take a staggered risk-based approach similar to the EU AI Act. California also recently passed an AI bill of its own, which introduces transparency obligations as regards training datasets used in generative AI systems.[562] Dozens, if not hundreds, of other AI regulatory initiatives are currently pending in other states.[563] China currently lacks comprehensive regulations for AI and has instead opted for regulating selected issues such as AI-generated content and algorithmic governance. China introduced its Interim Measures for generative AI in 2023, which will apply until it adopts more detailed rules as planned for in the coming years,[564] and has announced its intention to make AI a national priority.[565]

The road to international harmonization is very long indeed for digital sovereignty, and now also AI sovereignty. AI is clearly a technological product of national interest. As governments come to recognize that AI will determine their

[557] *National AI policies & strategies*, OECD.AI, https://oecd.ai/en/dashboards/overview [https://perma.cc/66WN-7LBE] (last visited Nov. 3, 2024).

[558] Draft Brazilian AI Regulations, *supra* note 379. *See supra* Section VII.A.2.

[559] Bill C-27, Artificial Intelligence and Data Act, 44th Parl., 1st Sess., 2022 (Can.).

[560] However, several presidential executive orders on AI have been passed. For example, Executive Order 14110 of Oct. 30, 2023 on "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," sets out detailed reporting requirements for AI developers based in the U.S. on planned activities with regard to dual-use foundation models and large-scale computing clusters in order to track their acquisition, development, or possession. Exec. Order No. 14110, 3 C.F.R. 657 (2024). This was justified on the basis of national security and cybersecurity. *Id.* at 657. The Executive Order is currently pending within the Department of Commerce for proposed regulations. *See* Taking Additional Steps to Address the National Emergency With Respect to Significant Malicious Cyber-Enabled Activities, 89 Fed. Reg. 5698 (proposed Jan. 29, 2024).

[561] S.B. 24-205, 2024 74th Gen. Assemb., Reg. Sess. (Colo. 2024).

[562] A.B. 2013, 2023-2024 Leg., Reg. Sess. (Cal. 2024).

[563] For an overview of failed, pending and enacted state AI-related regulations across the United States, see *Artificial Intelligence 2024 Legislation*, NCSL (Sept. 9, 2024), https://www.ncsl.org/technology-and-communication/artificial-intelligence-2024-legislation.

[564] In May 2024, the National Information Security Standardization Technical Committee released its new draft regulations, "Cybersecurity Technology – Basic Security Requirements for Generative Artificial Intelligence (AI) Service." Giulia Interesse, *China Releases New Draft Regulations on Generative AI*, CHINA BRIEFING (May 30, 2024), https://www.china-briefing.com/news/china-releases-new-draft-regulations-on-generative-ai/ [https://perma.cc/8DL2-85ZS].

[565] Graham Webster et al., *China's 'New Generation AI Development Plan' (2017)*, STAN. U.: DIGICHINA (Aug. 1, 2017), https://digichina.stanford.edu/work/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017/ [https://perma.cc/VH6Q-5HN8] ("AI has become a new focus of international competition. AI is a strategic technology that will lead in the future; the world's major developed countries are taking the development of AI as a major strategy to enhance national competitiveness and protect national security.")

future economic and industrial productivity, and even security and military capabilities,[566] they are shaping their policies and regulations to meet their own needs. Countries such as the United States, United Kingdom, France, Germany, India, China, Saudi Arabia and the United Arab Emirates have all pledged to support local AI development.[567] As a result, both companies and governments are intensifying efforts to advance their AI technologies domestically. Enacting regulations that govern AI deployed within a country's borders is just one means of control, even if a very important one. Indeed, this might explain why the EU was so concerned about ensuring "a level playing field among providers of general-purpose AI models where no provider should be able to gain a competitive advantage *in the Union market* by applying lower copyright standards than those provided in the Union."[568]

More recently, we have also witnessed export restrictions when it comes to AI. The United States is home to NVIDIA, whose GPUs and software have market dominance for powering the development and deployment of AI at scale.[569] In October 2022, the United States banned the export of NVIDIA's A100 and more advanced H100 chips to mainland China and Hong Kong.[570] A year later, in October 2023, this ban was extended to include the slower A800 and H800 chips, which had been specifically developed for the Chinese market.[571] Only recently did OpenAI also ban Chinese developers from accessing its service for building their custom GPTs.[572] In response, China has prioritized AI in its "Made in China 2025"

[566] For example, the National Security Commission on Artificial Intelligence in its 700-page report from 2021 concluded that: "We know adversaries are determined to turn AI capabilities against us. We know a competitor is determined to surpass us in AI leadership. We know AI is accelerating breakthroughs in a wide array of fields. We know that whoever translates AI developments into applications first will have the advantage. Now we must act." NAT'L SEC. COMM'N ON ARTIFICIAL INTELLIGENCE, FINAL REPORT 28 (2021), https://assets.foleon.com/eu-central-1/de-uploads-7e3kk3/48187/nscai_full_report_digital.04d6b124173c.pdf [https://perma.cc/QL82-KDKC].

[567] *See Welcome to the Era of AI Nationalism*, THE ECONOMIST (Jan. 9, 2024), https://www.economist.com/business/2024/01/01/welcome-to-the-era-of-ai-nationalism [https://perma.cc/JK3A-EG9E].

[568] EU AI Act, *supra* note 11, at Recital 106 (emphasis added).

[569] A recent report from CB Insights estimates that Nvidia has 95% of the GPU market for machine learning. Steve Salvius, *Nvidia's AI Dominance: How Long Will This Bull Run?*, SG ANALYTICS (Mar. 7, 2024), https://us.sganalytics.com/blog/nvidia-AI-dominance-how-long-will-this-bull-run/ [https://perma.cc/NVX3-DKMN].

[570] Tom Jowitt, *US Tightens AI Chip Export Restrictions To China*, SILICON (Apr. 1, 2024), https://www.silicon.co.uk/cloud/datacenter/us-tightens-ai-chip-export-restrictions-to-china-556794 [https://perma.cc/88CZ-FQM4].

[571] *Id.*

[572] Ben Jiang, *OpenAI's ban on Chinese access to ChatGPT to spur growth of local alternatives, experts say*, MYNEWS (June 26, 2024), https://www.scmp.com/tech/tech-war/article/3268124/openais-ban-chinese-access-chatgpt-spur-growth-local-alternatives-experts-say [https://perma.cc/2HJR-CLHQ].

initiative, aiming to reduce reliance on foreign technology and strengthen national security and to become the world's leading AI innovator by 2030.[573]

As governments become more involved in both publicly and privately held AI initiatives, national security interests that could be at stake add another layer of complexity that further weakens the position of rightholders. Taking Europe as an example, national security interests are not harmonized at the EU-level and are specifically excluded from the EU copyright regime.[574] Similarly, the EU AI Act specifically excludes Member States' competence concerning national security.[575] This recently came to the spotlight in Sweden, where it has been proposed that the Swedish Copyright Act shall specifically exclude from copyright protection the use of works in the interest of national security.[576]

The point is, therefore, that copyright extraterritoriality in the context of AI development is part of a much larger trend of maintaining digital sovereignty. It is a trend that is not just regulatory but highly political. If anything, the diverging and geopolitically sensitive context reinforces the need to adopt a more cautious approach when it comes to far-reaching extraterritorial measures. But importantly, it also suggests that the future of AI development may inevitably become a part of digital sovereignty in any event, as it appears that we may be slowly heading towards domestically built and launched AI models as a result of a more geographically fragmented regulatory landscape. Regulatory fragmentation that concerns the underlying development process, whether it is due to copyright, data privacy or AI regulations themselves, is highly problematic for AI developers and providers. In September 2024, more than 40 industry companies and associations, including Ericsson, Spotify and Meta, issued an open letter directed at EU

[573] Ulrich Jochheim, *China's ambitions in artificial intelligence*, EUROPEAN PARLIAMENT: THINK TANK (Sept. 2021), https://www.europarl.europa.eu/RegData/etudes/ATAG/2021/696206/EPRS_ATA(2021)696206_EN.pdf. *See also* Ian Burrows, *Made in China 2025: Xi Jinping's plan to turn China into the AI world leader*, CNAS (Oct. 5, 2018), https://www.cnas.org/press/in-the-news/made-in-china-2025-xi-jinpings-plan-to-turn-china-into-the-ai-world-leader [https://perma.cc/F3BV-UZUG].

[574] Infosoc Directive, *supra* note 97, at art. 9 ("This Directive shall be without prejudice to provisions concerning in particular . . . security . . ."). The European Court of Justice has confirmed in *Painer* that only Member States and their competent authorities, not private individuals or organizations, can invoke national security interests as an exception or limitation of copyright. Case C-145/10, Eva-Maria Painer v Standard Verlags GmbH, ECLI:EU:C:2011:798, ¶¶ 111-16 (Dec. 1, 2011).

[575] EU AI Act, *supra* note 11, at art. 2(3) ("This Regulation does not apply to areas outside the scope of Union law, and shall not, in any event, affect the competences of the Member States concerning national security, regardless of the type of entity entrusted by the Member States with carrying out tasks in relation to those competences."). *See also id.* at Recital 24.

[576] Utredningen om upphovsrättens inskränkningar [Committee on Copyright Limitations], Inskränkningarna i upphovsrätten [Limitations on Copyright], REGERINGSKANSLIET [GOV'T OFF. OF SWEDEN] 335-36, 424-25 (Jan 19, 2024), https://www.regeringen.se/contentassets/6b70735c6c05451c9728a4a2d987bf05/inskrankningarna-i-upphovsratten-sou-2024_4.pdf [https://perma.cc/973P-YZMR].

regulators, criticizing the "absence of consistent rules."[577] The same letter stated that "Europe can't afford to miss out on the widespread benefits from responsibly built open AI technologies that will accelerate economic growth and unlock progress in scientific research. For that we need harmonized, consistent, quick and clear decisions under EU data regulations that enable European data to be used in AI training for the benefit of Europeans."[578]

XII. CONCLUSION

The development of AI models revolves around data in vast quantities. The use of copyrighted materials when training models may be, as developers have admitted, a necessity for that purpose. Without sufficient source materials of high quality, including copyrighted works, we would not have witnessed the recent breakthroughs in recent years, particularly when it comes to generative AI models. AI is the fourth industrial revolution, which will enable economic, social and industrial benefits at a scale like we have never experienced before, and it is clear that it is here to stay.

Copyrighted materials have not lost their protection just because they become part of training datasets. The traditional rules of copyright infringement continue to apply, and unless exceptions and limitations exist that exempt such acts of reproduction from infringing, AI developers and providers may face substantial liability. So far, there is considerable divergence between jurisdictions in striking a fair balance between the interests of rightholders and AI stakeholders. The EU and Singapore stand out to be some of the most permissive up to now, whereas the copyright infringement position in the United States, Israel, and Japan will vary depending on the circumstances. Most other countries in the world, including for example the U.K., India, and Australia, presently have no exceptions and limitations for text and data mining, although that is likely to change in the near future.

In the absence of international harmonization for exceptions and limitations for text and data mining, there is a high likelihood that the same training activity will be infringing in some countries and not in others. The AI community is not blind to that risk. If copyright law severely restricts the development and deployment of AI, developers may decide to relocate their operations elsewhere, where the reproduction of training data is clearly not infringing. This Article has explained that there is a loophole in the international copyright system as it currently stands that would permit large-scale copying of training data in one country where this is not infringing. Once the training is done and the model is complete, developers could make the model available to customers in other countries, even if the same training activities would have been infringing *if* they had occurred there. This is possible due to the territoriality of copyright laws. If copyrighted materials are

[577] *An Open Letter: Europe needs regulatory certainty on AI*, EUNEEDSAI (Sept. 19, 2024), https://euneedsai.com/ [https://perma.cc/RVC2-5JQW].

[578] *Id.*

reproduced in training datasets, and if those acts of reproduction take place exclusively in one country where that is permitted, then that is the end of the story.

But this strict territorial approach is, of course, unsatisfactory to rightholders. It would be unfair to rightholders if they are denied legal recourse in their respective countries, which have taken the position that copyright protection should trump, merely because the reproduction activities in training the model took place somewhere else, which have taken the opposite position. What is arguably most threatening to rightholders is not so much lost licensing fees for the reproduction of their works during the training phase but the generated output capable of rendering once the AI model is complete. It is the completed model and its output, not the process behind it and its input, that could reduce the demand for original copyrighted works and potentially threaten authors' livelihood. This is, therefore, ultimately a product-by-process problem for rightholders, similar to what has long existed in the context of patent law. This raises a larger policy question about whether the proper focus in regulating AI from a copyright perspective should be on controlling the outputs (i.e., the AI-generated content itself) rather than the inputs (i.e., the data used to train AI models).

What remains so difficult for rightholders in this regard is that copyright law does not *per se* protect against the reduced demand of original, copyrighted works that could arise from AI-generated content. With the exception of "memorized" training data, the vast majority of AI-generated content will not directly infringe the training data as a direct consequence of the idea-expression dichotomy. They will not be infringing derivative works, nor they will be identical or substantially similar to the training data. Copyright laws are not designed to deal with new works and new reproductions that are statistically similar and created through a process using vast quantities of original works in fragmented parts. Because of that, dataset creators and model providers are also only expected to become secondarily liable for AI-generated works that directly infringe on copyright in exceptional cases. If there is no primary infringement, there can be no secondary infringement. It is ultimately a policy question whether the law should change in this regard.

In response to these issues, the EU has so far become the first to extraterritorially extend their copyright laws to text and data mining activities conducted in third countries. The message is that, regardless of where the training activities have occurred, if the completed model is later placed on the European market, then EU copyright law will apply retroactively to the training activity. While such an extraterritorial application benefits rightholders, and closes the loophole now present, it makes the situation significantly more complex for developers. If other regulators decide to follow the same path as the EU, as a means of maintaining digital sovereignty in the age of AI, then developers would be facing conflicting copyright laws targeting the same underlying activity. At least Brazil has so far indicated that it intends to do so, which also happened in the data privacy context. If this happens, then it would be particularly alarming because developing a model is typically only undertaken once due to the substantial investment required. It may be economically and technically unrealistic for developers to develop models on a country-by-country basis in order to satisfy diverging regulatory requirements. There is then a risk that developers are forced to choose between abandoning some domestic markets entirely or releasing less potent and powerful AI models using exclusively public domain and/or authorized datasets. Concurrently or alternatively, significantly increased development costs could drive up prices for AI models, making them less accessible to the public.

Both extremes present significant challenges. On the one hand, it is unjust to rightholders who benefit from copyright protection in a given country to see their rights become meaningless when training occurs elsewhere. On the other hand, imposing a burden on developers to curate training datasets that comply with varying and often conflicting copyright standards across different jurisdictions is equally problematic. The costs associated with developing AI models, especially those that require vast data quantities, are already substantial. Requiring developers to conduct separate training activities for different jurisdictions adds another layer of complexity and expense, unless the development conditions and economics for these models drastically change.

Extending copyright laws extraterritorially to regulate exclusively foreign conduct also has serious legal defects. This comes from the fact that we are concerned with a product-by-process problem. It is the reproduction of copyrighted materials at the process stage during training, which is suspect from a copyright point of view, but if that activity occurs entirely in another country, then it would interfere with foreign sovereignty to regulate it extraterritorially. If regulators wish to extend their copyright laws extraterritoriality to close the loophole that exists for training activities, and to do so in a way that is aligned with international norms, there may be much to learn from how patent law tackled similar issues, now several decades ago. International patent treaties harmonized to what extent patent laws can be applied extraterritorially to reach imported products derived from foreign manufacturing processes, while respecting that not all states would regulate the problem in the same way. This Article has discussed how this product-by-process problem in patent law compares with copyright law and AI models, when those models are the result of training activities conducted in foreign countries. If some states consider that rightholders should have a say in how their works are used for text and data mining purposes, then there is an urgent need for a similarly coordinated international effort in copyright law, which balances the interests of rightholders with the technical, regulatory and economic realities faced by developers. If history teaches us anything, it is that extraterritoriality is never a good idea for closing loopholes when there are markedly different substantive contexts, both in law and policy. The future success of AI will depend on collaborative policymaking in this regard, ensuring that both rightholders and developers can thrive while respecting that different countries will come to strike that balance differently.