THE HEART OF THE MATTER: COPYRIGHT, AI TRAINING, AND LLMS

by DANIEL GERVAIS,[*] HARALAMBOS MARMANIS,[**] NOAM SHEMTOV,[***] and CATHERINE ZALLER ROWLAND[****]

This article explores the intricate relationship between copyright law and artificial intelligence, including large language models (LLMs). It begins with a detailed technical overview of LLM functionality, including tokenization, word embeddings, and the various stages of LLM development. The authors then delve into the copyright implications of using protected works for both training LLMs and generating outputs. The paper argues that the training process likely constitutes prima facie copyright infringement through the reproduction and adaptation of copyrighted works. This occurs at multiple levels, including the creation of temporary copies during training and the embedding of numerical representations of training data within the LLM itself. The authors draw parallels between these AI processes and established legal concepts, such as the translation of computer code into executable formats. A thorough evaluation of potential copyright exceptions and limitations across various jurisdictions is presented. This includes an in-depth analysis of the fair use doctrine in the United States, with particular attention to how AI companies are attempting to draw parallels with previous cases like the Google Books cases. The paper also examines the text and data mining provisions in the European Union's Digital Single Market Directive and their applicability to AI training. The authors discuss emerging legislation, such as the EU AI Act, and its potential global impact on AI development and copyright law. They also address the complexities arising from the borderless nature of AI technology and the territorial limitations of copyright laws, which may lead to issues like forum shopping for AI training. Given the legal uncertainties surrounding AI and copyright, the paper proposes licensing as a key solution to balance innovation with copyright protection. The authors argue that global licensing agreements could harmonize practices and provide a consistent framework for responsible use of copyrighted works in AI development. The article concludes by reflecting on how copyright law has historically adapted to technological changes. However, it emphasizes that AI presents unprecedented challenges that may require novel legal and market-based approaches. The authors stress the importance of finding solutions that foster both technological

[*]Milton R Underwood Chair in Law, Vanderbilt University and Director, Vanderbilt Intellectual Property Program.
[**]EVP & CTO, Copyright Clearance Center, Inc.
[***]Chair of Intellectual Property and Technology Law at the Centre for Commercial Law, Queen Mary University of London and Director of the Queen Mary Intellectual Property Research Institute.
[****]General Counsel, Copyright Clearance Center, Inc. The authors would like to thank Ting Ting Lu for her valuable and multifaceted assistance with this article.

innovation and respect for intellectual property rights in the rapidly evolving AI landscape.

INTRODUCTION

Over the past year and a half, artificial intelligence (AI) has exploded onto the international scene. ChatGPT,[1] Claude,[2] Copilot,[3] and Gemini[4] are just a few examples of the exponential growth of AI systems that are based on large language

[1] *See Introducing ChatGPT*, CHATGPT (Nov. 30, 2022), https://openai.com/index/chatgpt/.
[2] *See* MAIN PAGE FOR CLAUDE, CLAUDE, https://claude.ai/ (last visited July 30, 2024).
[3] *See* SIGNIN PAGE FOR MICROSOFT COPILOT, MICROSOFT COPILOT, https://copilot.microsoft.com/ (last visited July 30, 2024).
[4] *See* MAIN PAGE FOR GEMINI, GEMINI, https://gemini.google.com/app (last visited July 30, 2024).

models (LLMs) and are dominating the market.[5] While AI is not new, public awareness of LLMs and their potential impact on our day-to-day lives is much more recent. One of the pressing questions now commonly debated is how AI technologies and copyright can coexist and lead to advancements both now and in the years to come. To foster a future that is both pro-copyright and pro-AI, it is essential to carefully navigate the intersection of these two critical domains, harnessing the power of both copyright and technology as engines of innovation.

In this article, we first explain in Part I the technology because how LLMs use copyrighted material is obviously relevant to the legal analysis of potential liability for copyright infringement. We then review in Part II the various aspects of copyright law that are implicated by LLMs. We do so first mostly from a US perspective but also adopt a comparative and international view of the matter, with the European Union providing the main point of comparison. Part III provides a n analysis of amendments to national copyright laws meant to deal specifically with the copyright aspects of LLMs, generally focusing on training data and not on outputs. A conclusion, offering a few thoughts on a constructive way forward, follows.

I. UNDERSTANDING LARGE LANGUAGE MODEL

In this Part, we review the way in which LLMs are trained and how they use human language, which many of the models learn by copying and processing copyrighted material such as books, newspaper, and journal articles.

A. The Role of Language

To understand the relationship between AI and copyright, we must first grasp the fundamentals of how AI systems that use large language models (LLMs) currently operate. At the core of human communication, whether verbal or written, lies the arrangement of words in sequences governed by the rules of syntax specific to a particular language, such as English, an indispensable part of the legal system.[6] While words themselves can be single or multi-character symbols, as seen in Chinese and Japanese, from a computer science perspective, human languages are classified as "natural languages."[7]

---

[5] *See* Patricio Cerda Mardini & Martyna Slawinska, *A Comparative Analysis of Leading Large Language Models*, MINDSDB (Mar. 11, 2024), https://mindsdb.com/blog/navigating-the-llm-landscape-a-comparative-analysis-of-leading-large-language-models.

[6] *See* Clark D. Cunningham et al., *Plain Meaning and Hard Cases*, 103 YALE L. J. 1561 (1994) (explaining that there are few areas of human activity where language matters more than law); *see also* Jim Chen, *Law as a Species of Language Acquisition*, 73 WASH. U. L. QUART. 1263 (1995) (comparing law to a "species of language acquisition.").

[7] *See* Javier Andrade et al., *Human-Centered Conceptualization and Natural Language*, *in* ENCYCLOPEDIA OF HUMAN COMPUTER INTERACTION (C. Ghaoui ed., 2006).

The study of processing and understanding natural languages is known as "Natural Language Processing" (NLP)[8] and "Natural Language Understanding,"[9] respectively, with the former term being more commonly used.[10] When it comes to textual content, books, articles, and other text-based artifacts are essentially compilations of word sequences.[11] However, computers, which operate using numbers, cannot directly comprehend the words of our language.[12] Therefore, an essential part of the software architecture for all systems that process text, including AI systems, is the representation of text using numerical values that enable the system to perform the required tasks.

B. Tokenization

Generative AI (GenAI) systems, which include ChatGPT-like models, utilize copies of copyrighted and public domain content such as books and articles for training.[13] This content is pivotal for training because the LLM's performance on a wide range of linguistic tasks benefits significantly from the use of these materials. The process of converting natural language text into a numerical representation involves several steps.[14] The first step is known as "tokenization," which can range from simple separation of words based on whitespace or other separator markings to more complex techniques like lemmatization and stemming, collectively referred to as "text normalization."[15] Through this process, the natural language text is transformed into a set of tokens, which are then used to form a "vocabulary" - a list of tokens, each with an associated numerical value. This vocabulary can then be used to represent any raw text input as a series of numbers.

---

[8] *See generally* DANIEL JURAFSKY & JAMES H. MARTIN, SPEECH AND LANGUAGE PROCESSING: AN INTRODUCTION TO NATURAL LANGUAGE PROCESSING, COMPUTATIONAL LINGUISTICS, AND SPEECH RECOGNITION (2009).

[9] *See id.*

[10] *See* Leila Amgoud & Henri Prade, *Can AI Models Capture Natural Language Argumentation*?, 6(3) INT'L J. COGNITIVE INFORMATICS & NAT. INTEL. 19 (2012).

[11] *See id.*

[12] *See id.*

[13] *See* Ryan Daws, *Openai: Copyrighted Data 'Impossible' To Avoid for AI Training*, AI NEWS (Jan. 9, 2024), https://www.artificialintelligence-news.com/2024/01/09/openai-copyrighted-data-impossible-avoid-for-ai-training/. This is a basis for the dozens of US lawsuits alleging copyright infringement that are pending as of this writing against providers of LLMs, including OpenAI. *See infra* note 34.

[14] *See* ANNAMALAI CHOCKALINGAM ET AL., A BEGINNER'S GUIDE TO LARGE LANGUAGE MODELS: PART 1 (2023) (ebook).

[15] *See* DAVID FOSTER, GENERATIVE DEEP LEARNING: TEACHING MACHINES TO PAINT, WRITE, COMPOSE, AND PLAY 146-49 (2nd ed. 2023); *see also* Rene Y. Choi et al., *Introduction to Machine Learning, Neural Networks, and Deep Learning*, 9 TRANSLATIONAL VISION SCI. & TECH. 14 (2020).

The modern option for tokenizers is "subword tokenization," which produces sets of tokens that are smaller than words.[16] It should be noted that OpenAI and Azure OpenAI (as well as many others) use a subword tokenization method called "Byte-Pair Encoding (BPE)" for their Generative Pretrained Transformer (GPT)-based models.[17] BPE is a method that combines the most frequent character pairs into a single token, until a certain number of tokens or a vocabulary size is reached. The larger the vocabulary size, the more diverse and expressive the texts that the model can generate. A distinct advantage of subword tokenizers is their ability to handle out-of-vocabulary words.

In sum, tokenization allows us to map raw text onto a set of numbers, and a set of numbers back into text. That mapping is lossless. If you encode a string with your tokenizer and then decode what you get from the encoder, you will get back exactly the text that you used as the input. However, tokenization is the first step toward a numerical representation of the sequences of words that occur in text, and while it is necessary, it is not sufficient, especially for the kind of AI applications that have taken the world by storm in recent years. For these applications, we need so-called "dense" representations.[18]

C. Embeddings

Enter the world of word embeddings, which provide the representation of a word as a high-dimensional vector.[19] You can think of vectors as rows of numbers (or columns for that matter) in an Excel spreadsheet. However, the dimensionality of these vectors is much smaller than the size of the vocabulary, and therefore much smaller than the dimensionality of the vector representations that we can build based directly on the vocabulary.[20] They are also dense (instead of sparse), consist of real values (instead of integers), and they turn out to be far more effective and efficient for all NLP tasks. It is important to note that word embeddings can and are used in AI systems that process text, whether an LLM is used or not.

---

[16] *See* Thinh Hung Truong et al., Revisiting Subword Tokenization: A Case Study on Affixal Negation in Large Language Models (Apr. 4, 2024) (unpublished manuscript), https://arxiv.org/abs/2404.02421.

[17] *See* Alec Radford et al., Improving Language Understanding by Generative Pre-Training (June 10, 2018) (unpublished manuscript), https://paperswithcode.com/paper/improving-language-understanding-by; *see also Byte-Pair Encoding*, HUGGING FACE, https://huggingface.co/learn/nlp-course/en/chapter6/5 (last visited July 30, 2024).

[18] *See* Tomas Mikolov, Wen-tau Yih & Geoffrey Zweig, *Linguistic Regularities in Continuous Space Word Representations, in* PROC. 2013 CONF. OF THE NORTH AM. CHAP. OF THE ASS'N FOR COMPUT. LINGUISTICS: HUMAN LANG. TECHS. 746-51 (Lucy Vanderwende et al. eds., 2013).

[19] *See* Tomas Mikolov, Kai Chen, Greg Corrado & Jeffrey Dean, Efficient Estimation of Word Representations in Vector Space (Sept. 7, 2013) (unpublished manuscript), https://arxiv.org/abs/1301.3781.

[20] *See id.*

The important part, from a copyright perspective, is that modern AI systems create and store contextual word embeddings that capture the relationship of words in long sequences.[21] The Transformer architecture is all about attention, as the groundbreaking paper described it: "Attention is all you need."[22] Hence, arguing that there are no copies after training because there are only "parameters" (probably referring to the weights of the LLM) is incorrect because these "parameters" actually define the probability distribution over enormously large vector spaces[23]. If you take two words from a vocabulary of 100 words, then there are only 9900 possible combinations, but if you take 1000 words from a vocabulary of 100,000 words, the number of possible combinations is astronomical, and the only combinations that would have non-negligible weights (i.e., non-zero probability of occurrence) would be the ones that were observed during training. Thus, in a number of cases, this will result in a lossless reproduction of materials that were used during training, causing what is sometimes referred to as "memorization."[24] The more unique the material, the more likely it is for that to happen. ChatGPT (in its GPT-3.5 incarnation) used to regurgitate Dr. Seuss by simply prompting it with the text "Oh, the places you'll go." That was no accident.[25]

At this point, we should note that an analogy can be drawn between AI systems that process copyrighted works written in human language, on one hand, and the special programs that are used to translate any human readable code into an executable representation of that code, on the other hand. When humans write computer programs, we use special languages (called "programming languages") such as C, Java, Python, and so on.[26] For some of these languages (e.g., C) the programs (which can be easily understood by a human who can speak that special language) are translated into "executable files" by a special program called the

---

[21] *See* Francesca Incitti, Federico Urli & Lauro Snidaro, *Beyond word embeddings: A survey*, 89 INFO. FUSION 418-36 (2023).

[22] Ashish Vaswani et al., Attention Is All You Need (June 19, 2017) (unpublished manuscript), https://arxiv.org/pdf/1706.03762v2.pdf.

[23] *See* Harry Surden, *ChatGPT, AI Large Language Models, and Law*, 92 FORDHAM L. REV. 1941, 1959 (2024) ("GPT systems are just a series of billions or trillions of numbers, known as parameters. These parameters are what guides the system's predictions to select one word ("Paris") versus others ("tree" or "zebra") among its over 50,000-word vocabulary."). [24] *See* A. Feder Cooper & James Grimmelmann, *The Files are in the Computer: Copyright, Memorization, and Generative AI*, CHI.-KENT L. REV. (forthcoming 2024) (manuscript at 12), https://arxiv.org/pdf/2404.12590 (Defining "memorization" as when "an exact or nearly-exact copy of a piece of training data can be reconstructed by examining the model through any means (not necessarily through prompting).").

[25] *See* André V. Duarte et al., DE-COP: Detecting Copyrighted Content in Language Models Training Data (Feb. 15, 2024) (unpublished manuscript), https://arxiv.org/pdf/2402.09910 (explaining how memorization can be detected).

[26] *See generally Programming language*, WIKIPEDIA, https://en.wikipedia.org/wiki/Programming_language (last visited July 16, 2024).

"compiler."[27] The executable files can be "understood" (executed) directly by the computer. For other languages, the program is translated into a special intermediate form that another special program (called the "interpreter") will use to turn into a form that can be understood by the computer. In both cases, we have the original copyrighted work translated into another form that is not directly understandable by humans. If we tried to reverse engineer the executable versions of a program you will not obtain the exact expression of the original program, but no one would argue that the content of the original is absent from that machine-understandable version of the program. Analogously, no one should argue that the content of the original text is absent from the LLM representation.

In summary, word embeddings capture, in a dense vector representation, the meaning of a word in the context of the words that surround it, as found in the text that is used during training. So, in essence, word embeddings are a mathematical construct that can efficiently capture the meaning of words based on the various contexts (i.e., word sequences) in which a word can be found in. This has been known since the 1950s as the distributional hypothesis.[28] Word embeddings are fundamental blocks during the construction and operation of LLMs as discussed in the next section.

D. The Stages of LLMs

Large language models go through various stages of maturity, from inception to production. We can identify four major stages in their training evolution: (1) pretraining, (2) supervised fine-tuning, (3) reward modeling, and (4) reinforcement learning.[28] Everything after pretraining is essentially fine-tuning. The focus in pretraining is on quantity whereas the focus during fine-tuning is on the quality of the material used. However, in all fine-tuning stages and all downstream applications, the representation of the content stored in an LLM is leveraged because that representation is permanent and always active during the operation of a LLM.

Perhaps the most interesting area of fine-tuning is for Span-based applications.[29] That case involves generating, and operating with, representations of contiguous sequences of tokens. Examples of applications include named entity recognition (NER), coreference resolution, Q&A, syntactic parsing, and so on.[30] In the context of the latter applications, and in light of the copyright issues, one should consider the morphological characteristics of what an LLM based system produces in juxtaposition with its input. In other words, the question is how

---

[27] *See Compiler*, WIKIPEDIA, https://en.wikipedia.org/wiki/Compiler (last visited July 16, 2024).

[28] *See* Microsoft Developer, *State of GPT | BRK216HFS*, YOUTUBE (May 25, 2023), https://www.youtube.com/watch?v=bZQun8Y4L2A.

[29] *See* Zhengao Jiang et al., *Generalizing Natural Language Analysis through Span-relation Representations*, *in* PROC. 58TH ANN. MEETING ASS'N FOR COMPUT. LINGUISTICS 2120-33 (2020).

[30] *See id.*

similar the text of an answer would be in relation to the text of the content that was used during pre-training or fine-tuning.

LLMs make copies of the documents on which they are trained and this copying takes various forms, and, as a result, with appropriate prompting applications that use the LLMs are able to reproduce original works. The internal representations of the text on which they are trained, in purpose-built vector spaces, are very different in nature from those used in traditional search applications based on indexing because the latter systems consider only the relevance of a given query to the indexed terms of each document, they cannot recreate the indexed documents based on their internal representations -- the only way to do this is to actually store a copy of the original text.[31]

It should also be noted that the various forms of copying involve copies that are permanent in nature, such as the initial copies in the training set or the internal representations of the processed text, and transient in nature such as copies made to support the transfer of information between different parts of an AI system or copies related to the output generated during the use of an AI system in what is typically called a "user session."

Our discussion, so far, has been with reference to AI systems that are trained with text. However, modern AI systems are multimodal, which means that they are able to train on and generate text, images, audio, as well as video. The most prevalent models for non-textual generation are the so-called diffusion models, a class of probabilistic generative models that progressively diffuse training data with injected noise and then learn to reverse this process to generate new data from the noise. All these models end up being specific cases of a generalized stochastic differential equation.[32] We cannot provide a detailed account of these models since they require a certain level of mathematical expertise.[33] The important element of their character is the similarity that they bear with LLMs in that the outcome of their learning process results in a sophisticated permanent copy of the input.

II. COPYRIGHT LAW ASPECTS OF LARGE LANGUAGE MODELS

LLMs are using large quantities of textual works to train and produce the best results, with most LLMs relying on vast amounts of copyrighted works.[34]

---

[31] *See* STEPHEN BUTTCHER ET AL., INFORMATION RETRIEVAL: IMPLEMENTING AND EVALUATING SEARCH ENGINES (2016).

[32] *See Diffusion model*, WIKIPEDIA, https://en.wikipedia.org/wiki/Diffusion_model (last visited July 16, 2024).

[33] For a detailed account *see* LING YANG ET AL., DIFFUSION MODELS: A COMPREHENSIVE SURVEY OF METHODS AND APPLICATIONS (2024).

[34] At least according to the multiple pending lawsuits against LLM providers alleging copyright infringement. *See, e.g.,* Doe 1 v. Github, Inc., No. 4:22-cv-06823, 2023 WL 3449131 (N.D. Cal. May 11, 2023); Andersen v. Stability AI Ltd., No 3:23-cv-00201, 2023 WL 7132064 (N.D. Cal. Oct. 30, 2023); Getty images (US), Inc. v. Stability AI, Inc., No.

Copyright law protects all original works—those that are independently created and meet a low standard of creativity.[35] Copyright law gives copyright owners the exclusive right of reproduction (copying), among other rights.[36] When a third party, such as an LLM developer, makes copies of massive amounts of copyrighted works without permission, a range of copyright infringement issues arise.[37]

Some works used for trainng arenot or are no longer protected by copyright, like those for which the copyright term has expired and U.S. government works.[38] However, according to many of the pending lawsuits, a number of LLMs were trained on copyrighted material, making copies of the material before, during, and possibly after training and fine-tuning processes.[39]

A. Copyright and AI in Historical Perspective

The concept of copyright has its roots in the early days of print culture and has evolved dramatically over time in response to technological advances.[40] The Statute of Anne, passed in Great Britain in 1709, is often cited as the first major

---

1:23-cv-00135 (D. Del. Feb.3,2023); *In re* OpenAI ChatGPT Litigation, No. 3:23-cv-03223, 2024 WL 2044625 (N.D. Cal. May 7, 2024); Tremblay v. OpenAI, Inc., No. 3:23-cv-03233, 2024 WL 557720 (N.D. Cal. Feb. 12, 2024); Kadrey v. Meta Platforms, Inc., No. 3:23-cv-03417, 2023 WL 10673221 (N.D. Cal. Dec. 1, 2023); J. L. v. Alphabet, Inc., No. 3:23-cv-03440, 2024 WL 3282528 (N.D. Cal. June 6, 2024); Thaler v. Perlmutter, 687 F. Supp. 3d 140 (D.D.C. 2023); Authors Guild v. OpenAI Inc., No. 1:23-cv-08292 (S.D.N.Y. Sept. 18, 2023); Alter v. OpenAI Inc., No. 1:23-cv-10211 (S.D.N.Y. Nov. 21, 2023); Huckabee v. Meta Platforms, Inc., No. 3:23-cv-06663 (N.D. Cal. Oct. 17, 2023); Concord Music Group, Inc. v. Anthropic PBC, No. 3:23-cv-01092 (M.D. Tenn. Oct. 18, 2023); N.Y. Times Co. v. Microsoft Corp., No. 1:23-cv-11195 (S.D.N.Y. Dec. 27, 2023); The Intercept Media, Inc. v. OpenAI, Inc., No. 1:24-cv-01515 (S.D.N.Y. Feb. 28, 2024); Raw Story Media, Inc. v. OpenAI Inc., No. 1:24-cv-01514 (S.D.N.Y. Feb. 28, 2024); Nazemian v. NVIDlA Corp., No. 4:24-cv-01454 (N.D. Cal. Mar. 8, 2024); Zhang v. Google LLC, No. 3:24-cv-02531 (N.D. Cal. Apr. 26, 2024); Daily News LP v. Microsoft Corp., No. 1:24-cv-03285 (S.D.N.Y. Apr. 30, 2024); Dubus v. NVIDIA Corp., No 3:24-cv-02655 (N.D. Cal. May 2, 2024); Makkai v. Databricks, Inc., No. 4:24-cv-02653 (N.D. Cal. May 2, 2024); UMG Recordings, Inc. v Suo, Inc., No. 1:24-cv-04777 (S.D.N.Y. Jun. 24, 2024); The Ctr. For Investigative Rep. v. OpenAI, Inc., No 1:24-cv-04872 (S.D.N.Y. Jun. 27, 2024); Bartz v. Anthropic PBC, No. No. 3:24-cv-05417 (N.D. Cal. Aug. 19, 2024) [hereinafter, collectively, the US Cases].

[35] 17 U.S.C. § 101; Berne Convention for the Protection of Literary and Artistic Works art. 2(1), Sept. 9, 1886, as revised at Paris on July 24, 1971 and amended in 1979, S. Treaty Doc. No. 99-27 (1986) [hereinafter Berne Convention].

[36] 17 U.S.C. § 106.

[37] Feist Publications, Inc. v. Rural Tel. Serv. Co., 499 U.S. 340, 361 (1991).

[38] 17 U.S.C. §§ 105, 302(a).

[39] *See supra* note 34.

[40] DANIEL J. GERVAIS, *Chapter 1: Copyright in common law jurisdictions*, *in* (RE)STRUCTURING COPYRIGHT: A COMPREHENSIVE PATH TO INTERNATIONAL COPYRIGHT REFORM (rev. ed., 2019).

copyright law, signaling the transition from royal grants to a system designed to encourage creativity while allowing the reproduction and distribution of works.[41] It provided protection for authors and book publishers and fostered a sustainable market for literature.[42]

As technology advanced, new forms of expression, such as music and theater, began to flourish. Laws were expanded to facilitate live public performances and ensure compensation for authors of musical works and plays.[43] The invention of the player piano in the late 19th century—"mechanizing" music reproduction—necessitated an update of copyright laws to include mechanical reproduction rights.[44]

The advent of radio broadcasting in the early 20th century prompted the extension of performance rights to those broadcasts.[45] Similarly, with the birth of cinema at the turn of the 20th century, motion pictures were recognized as a new category of copyrighted works.[46] It is actually possible to follow the timeline of adaptations of the copyright framework by tracking the frequent amendments to U.S. copyright law and international copyright instruments. The most important copyright treaty is the Berne Convention for the Protection of Literary and Artistic Works.[47] It was first adopted in 1886 and as of June 2024, it has 181 member states, making it the de facto basic global reference point for copyright. It was modified (typically by the adoption of new versions or "Acts") in 1896, 1908, 1928, 1948, 1967 and 1971.[48] New categories of works have been added to reflect changing technology (such as motion pictures, added in 1928) or new ways of exploiting and administering rights (such as the right in sound recordings, modified in 1948, or cable retransmissions in 1967).[49]

Each revision aimed to modernize the Convention in the light of technological, cultural, and legal developments around the world, broadening the scope of works covered and strengthening the rights of authors and other copyright holders. The Paris Act of 1971 is the most recent and current act under the Berne Convention.[50] After 1971, copyright protection was extended to computer software as a literary work, which was confirmed globally by the 1994

---

[41] MARK ROSE, AUTHORS AND OWNERS: THE INVENTION OF COPYRIGHT (1993).
[42] *Id.*
[43] *See* PAUL GOLDSTEIN, GOLDSTEIN ON COPYRIGHT 1.13.2 (3rd ed. 2023).
[44] *See id.* at 7.2.
[45] *See id.* at 7.3.
[46] *See id.* at 2.12.
[47] Berne Convention, *supra* note 35. As of June 2024, it had 181 member States. *See Berne Convention Contracting Parties*, WIPO LEX, https://www.wipo.int/wipolex/en/treaties/ShowResults?search_what=C&treaty_id=15 (last visited July 30, 2024).
[48] WIPO, BERNE CONVENTION CENTENARY 1886-1986 (1986).
[49] 1 JANE GINSBURG & SAM RICKETSON, INTERNATIONAL COPYRIGHT AND NEIGHBOURING RIGHTS: THE BERNE CONVENTION AND BEYOND ¶ 3.01 (3rd ed. 2022).
[50] *See id.*

Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS) administered by the World Trade Organization (WTO).[51]

The emergence of the World Wide Web in the 1990s created unprecedented challenges and opportunities for copyright holders. To address these, two major treaties were adopted in 1996 under the auspices of the World Intellectual Property Organization (WIPO): the WIPO Copyright Treaty (WCT) and the WIPO Performances and Phonograms Treaty (WPPT).[52] These treaties recognized the "making available" right in the digital environment and addressed the need to protect rights management information online.[53] The WCT also contains an "agreed statement" according to which "storage of a protected work in digital form" is a reproduction and noting that the reproduction and exceptions thereto "fully apply in the digital environment," a phrase whose meaning as a matter of legal interpretation is not pellucidly clear.[54] What is clear, however, as a matter of international law is that copies stored for more than transitory duration in digital form are reproductions that must be either authorized or covered by an exception.[55]

The timeline of copyright legislation shows a clear pattern: the introduction of major new technologies has often been followed by the creation of new exclusive rights (e.g., broadcasting rights) and/or compensation mechanisms (e.g., cable retransmission fees), as well as limitations and exceptions (e.g., for parody or criticism) that reflect a balance between the interests of copyright holders and those of the broader public.[56] Courts have also tried to reflect a balance when interpreting the statute. We see examples in opposite directions in opinions of the Supreme Court in *Sony* and *Grokster*.[57] In the former case, the court moved from an apparent position of significant skepticism at oral argument to an affirmation of fair use for the sale of home video recording devices (VCRs) as a dual-use technology capable of both infringing and substantial non-infringing.[58] In the

---

[51] Agreement on Trade-Related Aspects of Intellectual Property Rights art 10.1, Apr. 15, 1994, 1869 U.N.T.S. 299 [hereinafter TRIPS Agreement] ("Computer programs, whether in source or object code, shall be protected as literary works under the Berne Convention (1971).").

[52] WIPO Copyright Treaty, Dec. 20, 1996, 2186 U.N.T.S. 121 [hereinafter WCT]; WIPO Performances and Phonograms Treaty, Dec. 20, 1996, 2186 U.N.T.S. 203 [hereinafter WPPT].

[53] *See* WCT, *supra* note 52, arts. 6.1, 8; WPPT, *supra* note 52, arts. 8.1,10, 12.1,14, 19.2.

[54] Among the questions the Agreed Statement to article 1(4) does not answer are: does it bind all parties to the Berne Convention, even those that neither signed nor adhered to the WCT? Does it apply to the interpretation of the TRIPS Agreement? For a discussion, *see* GINSBURG & RICKETSON, *supra* note 49, ¶¶ 11.79-11.88.

[55] *See id.*

[56] GERVAIS, *supra* note 40, 207-215.

[57] Sony Corp. of Am. v. Universal City Studios, Inc., 464 U.S. 417 (1984); Metro-Goldwyn-Mayer Studios Inc. v. Grokster, Ltd., 545 U.S. 913 (2005).

[58] *See* Jessica Litman, *The Story of* Sony v. Universal Studios*: Mary Poppins Meets the Boston Strangler, in* INTELLECTUAL PROPERTY STORIES 358, 366-368 (J. C. Ginsburg & R. C. Dreyfuss eds., 2006).

latter case, although peer-to-peer file sharing was also a dual use technology, its "promotion" as an infringement tool led the court to find Grokster secondarily liable under a doctrine of inducement borrowed at least in part from patent law.[59]

Today, we stand on the brink of another crucial era in which technologies such as AI can absorb and process the creative works of humans to autonomously produce competitive content.[60] The stakes for creators, industries, and society as a whole are immense, as the fundamental nature of creativity and the way we value human artistry may be about to change. It would be strange if, for perhaps the first time, such a significant change was not met with adequate adaptation of the international legal framework.[61]

B. Infringement Analysis

Like with other advancements, copyright concepts apply to AI and help set the stage for how we can use both technology and the law to promote innovation. It is clear that AI is built on a foundation of immense works of authorship, many of which are protected by copyright.[62] When copyrighted works are used, AI systems typically make copies of the works to train and power AI outputs.[63] In LLMs, tokenization, sub-word tokenization, and the creation and storage of contextual word embeddings that delineate the relationships between words in extensive sequences can implicate two distinct exclusive copyright rights: the right of reproduction (essentially, copying), and the right of adaptation.[64] The reproduction right essentially gives copyright owners the exclusive right to make

---

[59] *See Sony Corp.,* 464 U.S. at 442 ("Accordingly, the sale of copying equipment, like the sale of other articles of commerce, does not constitute contributory infringement if the product is widely used for legitimate, unobjectionable purposes. Indeed, it need merely be capable of substantial noninfringing uses."); *Grokster*, 545 U.S. at 948 ("Even if the absolute number of noninfringing files copied using the Grokster and StreamCast software is large, it does not follow that the products are therefore put to substantial noninfringing uses and are thus immune from liability.").

[60] *See* Daniel J. Gervais, *The Machine As Author*, 105 IOWA L. REV. 2053, 2057 (2020):

> [M]achines are increasingly good at emulating humans and laying siege to what has been a strictly human outpost: intellectual creativity. At this juncture, we cannot know with certainty how high machines will reach on the creativity ladder when compared to, or measured against, their human counterparts, but we do know this: They are far enough already to force us to ask a genuinely hard and complex question, one that intellectual property scholars and courts will need to answer.

[61] One of us has proposed to open a discussion at WIPO on the revision of the Berne Convention, which, like the work on a possible protocol to the Convention in the 1980s and early 1990s, could lead to a new instrument (in that case, the WCT). *See generally* GERVAIS, *supra* note 40 at Appendix. On the possible protocol, *see* GINSBURG & RICKETSON, *supra* note 49 ¶¶4.15-4.18.

[62] For example, *see* HOUSE OF LORDS COMMUNICATIONS AND DIGITAL SELECT COMMITTEE, OPENAI — WRITTEN EVIDENCE (2024), https://committees.parliament.uk/writtenevidence/126981/pdf/.

[63] *See infra* Part II.B.

[64] 17 U.S.C. §§ 106(1)-106(2).

copies of their work or to authorize others to do so.[65] Article 12 of the Berne Convention, provides that "Authors of literary or artistic works shall enjoy the exclusive right of authorizing adaptations, arrangements, and other alterations of their works."[66] Adaptation is broadly recognized as the process of transforming a work into a different form of expression, moving beyond simple reproduction, such as the adaptation of a novel into a film. In the United States, there is also the specific right to create derivative works—works based on the original work—that are enjoyed by the copyright owner.[67] Other nations tend to use the Berne Convention language and refer to adaptation and translation.[68]

If someone other than the copyright owner reproduces, adapts, or makes derivative works of a copyrighted work without permission, the copyright owner can make a claim of infringement, which would be subject to copyright law's set of exceptions and limitations.[69] In the case of LLMs, it is first necessary to determine whether there is unauthorized copying, adapting, or preparation of derivative works during the creation of training datasets, as well as during the actual training process.[70] Specifically, does the comprehensive training process involve the unauthorized replication of works, and does a transformer model, once trained, retain copies or unauthorized adaptations of protected works?

1. Inputs

As discussed below, in instances where an AI system replicates a piece of protected content during its training phase, this is likely to be viewed as prima facie infringement of the copyright holder's exclusive right to reproduce.[71] However, is the legal position less straightforward when considering "transient"

---

[65] 17 U.S.C § 106(1); RESTATEMENT OF COPYRIGHT § 56 (Am. L. Inst., Tentative Draft No. 3, 2022).

[66] Berne Convention, *supra* note 35, art. 12.

[67] *See* GINSBURG & RICKETSON, *supra* note 49, ¶¶ 11.30 - 11.41 (discussing the rights of reproduction and adaptation in the Convention).

[68] *See, e.g.*, German Act on Copyright and Related Rights (Urheberrechtsgesetz – UrhG), Section 3 – Adaptations (Copyright Act of 9 September 1965 (Federal Law Gazette I, p. 1273), as last amended by Article 25 of the Act of 23 June 2021 (Federal Law Gazette I, p. 1858); UK Copyright, Designs and Patents Act 1988, Section 21 – adaptation.

[69] *See* 17 U.S.C. § 501(a) ("Anyone who violates any of the exclusive rights of the copyright owner as provided by sections 106 through 122 . . . is an infringer of the copyright or right of the author, as the case may be."). The reference to "sections 106 through 122" includes sections 107 to 122 of the Act, which provide a set of exceptions and limitations on the rights granted to copyright owners and authors.

[70] This is part of the plaintiff's necessary prima facie case. *See Sony,* 464 U.S. at 432-33 (defining an infringer as"anyone who trespasses into his exclusive domain by using or authorizing the use of the copyrighted work in one of the ways set forth in the statute.").

[71] This point is presently being argued in numerous lawsuits around the world. *See supra* note 35 (US Cases); Getty Images (US) Inc & Ors v Stability AI Ltd [2023] EWHC 3090 (Ch) United Kingdom); Robert Kneschke v. LAION e. B, [27.04.2023] Hamburg *remuneration obligations in Europe*, GEMA, Nov. 13, 2024, https://www.gema.de/en/w/gema-files-lawsuit-against-openai (Germany).

duplications or when the training process utilizes only sections of the copyrighted content? Does the transient or "imperfect" nature of these copies alter the conclusion regarding infringement?

In scenarios involving the training of AI models, the answer is likely to be negative as the cases have long considered imperfect or incomplete and temporary copies as potentially infringing.[72] The legal stance is relatively clear around the world. For instance, under Article 2 of the EU's Information Society Directive, the exclusive right of reproduction encompasses "direct or indirect, temporary or permanent reproduction [...]."[73] This is subject to Article 5(1) of the Directive, which provides for an exception in the case of temporary reproductions under a strictly defined set of conditions.[74] This conclusion is similarly upheld in the case of the United States.[75] The US Copyright Office has opined that "Congress intended the copyright owner's exclusive right to extend to all reproductions from which economic value can be derived. The economic value derived from a reproduction lies in the ability to copy, perceive, or communicate it. Unless a reproduction manifests itself so fleetingly that it cannot be copied, perceived or communicated, the making of that copy should fall within the scope of the copyright owner's exclusive rights […] this would cover the temporary copies that are made in [random access memory ("RAM")] in the course of using works on computers and computer networks."[76] This distinction is highly pertinent in the context of AI training.

The copies made before and during training are generally not transient in nature as a matter of copyright law. Thus, it is clear that the reproduction made during the training process is sufficiently enduring for the AI model to "perceive" it, hence enabling "tokenization" to occur and consequently the derivation of economic value as referred to by the US Copyright Office.[77] This issue was settled by courts in the past in the context of computer programs. For example, the status of transient copies under copyright law was argued in the United States case MAI Systems Corp. v. Peak Computer, Inc.[78] Inter alia, this 1993 case dealt with whether loading a software program into a computer's RAM created a copy that could be considered a violation of copyright law. The court concluded that the

---

[72] Under US law, the plaintiff's burden is to prove the copying of "constituent elements of the work that are original." Feist Publ'ns, Inc. v. Rural Tel. Serv. Co., 499 U.S. 340, 361 (1991).

[73] Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the Harmonisation of Certain Aspects of Copyright and Related Rights in the Information Society, art. 2, 2001 O.J. (L 167) 22.6.2001.

[74] *Id.*

[75] *See* H.R. Rep. No. 94-1476, at 61 (1976), at 52-53 ("[T]he definition of 'fixation' would exclude from the concept purely evanescent or transient reproductions such as those projected briefly on a screen, shown electronically on a television or other cathode-ray tube, or captured momentarily in the 'memory' of a computer.").

[76] U.S. COPYRIGHT OFF., DIGITAL MILLENNIUM COPYRIGHT ACT §104 STUDY 111-12 (2001).

[77] *Id.*

[78] MAI Systems Corp. v. Peak Computer, Inc., 991 F.2d 511 (9th Cir. 1993).

creation of these transient copies in RAM did constitute a copy under the Copyright Act, thus implicating the reproduction right.[79]

While prima facie infringing copies are generated during the training process, such copies are not necessarily retained in an exact form once a transformer model has been fully trained on the basis of the copies. What the AI models retain post-training are contextual word embeddings that encapsulate the relationships between words in lengthy sequences.[80] The capacity of such systems to reproduce verbatim copies of protected text used as training material, sometimes producing exact or nearly exact copies that are thousands of words long if not more, could be attributed to the fact that these AI systems retain copies, adaptations, or derivative works, stored within the AI systems in specific numerical formats.[81] The fact is that these numerical representations could often be "worked backwards" to recreate a precise and complete version of the original content used as training material. For example, it has been unequivocally demonstrated that by performing data extraction attacks, it is possible to recover individual training data examples.[82] Consequently, following training, many AI models incorporate representations of the training data. A key question, therefore, is whether these representations can be considered infringing copies under copyright law.

Under copyright law, similarity between the original work and the allegedly infringing work, coupled with access by the alleged infringer to the original work, may give rise to a legal presumption of copying.[83] Applying this principle in an AI context leads to a similar conclusion: when the output is indeed highly similar if not identical, and the AI model had "access" to the original in the sense of being trained on it, copying may be presumed.[84] But how exactly? At the time such

---

[79] It is important to note that the US doctrine of fair use may be highly relevant when considering reproduction resulting from use in certain technological environments, as discussed below. However, the present analysis addresses prima facie infringement before considering any copyright exceptions.

[80] *See* CHOCKALINGAM et al., *supra* note 14.

[81] The example given above that ChatGPT (in its GPT-3.5 incarnation) used to regurgitate Dr. Seuss by simply prompting it with the text "Oh, the places you'll go" illustrates this point. *See supra*, I(C).

[82] For a comprehensive discussion on the presence of permanent copies' fragments of training material within generative AI models, based on experiments with a large range of language models, *see* ANTONIA KARAMOLEGKOU, ET AL., COPYRIGHT VIOLATIONS AND LARGE LANGUAGE MODELS (2023), arXiv:2310.13771v1.

[83] *See* Three Boys Music Corp. v. Bolton, 212 F.3d 477, 486 (9th Cir. 2000) ("By establishing reasonable access and substantial similarity, a copyright plaintiff creates a presumption of copying." (citing Granite Music Corp. v. United Artists Corp., 532 F.2d 718, 721 (9th Cir. 1976)); *see also* Herbert Rosenthal Jewelry Corp. v. Kalpakian, 446 F.2d 738, 741 (9th Cir. 1971) ("It is true that defendants had access to plaintiff's [copyrighted] pin and that there is an obvious similarity between plaintiff's pin and those of defendants. These two facts constitute strong circumstantial evidence of copying.").

[84] A plaintiff may prove copying by showing the existence of striking similarity between the allegedly infringing material and the plaintiff's work. *See* La Resolana Architects, PA

outputs are generated, transient copies created during the training process have likely been discarded.[85] Such copying is therefore taking place in relation to the numerical representations that are permanently embedded within the AI model. Hence, the numerical representations of the training data that are permanently embedded in LLMs may be considered as copies or adaptations of the original training material. For instance, consider the example of ChatGPT generating text similar to Dr. Seuss's "Oh, the Places You'll Go" when prompted with that phrase.[86] This illustrates a key point: during training, ChatGPT was exposed to "Oh, the Places You'll Go" along with numerous other works. As a result, the AI model created and stored numerical representations of contextual word embeddings that capture the relationships between words in the training data across long sequences. These representations define the probability distribution of text over extensive vector spaces. After the training phase, any transient copies of the training data were likely discarded. However, the ability of ChatGPT to, at a later time, to produce text reminiscent of Dr. Seuss based on these numerical representations makes a compelling case for considering such representations as infringing copies under copyright law

The analogy of converting computer programming language content into executable code also provides a useful comparison.[87] Even though the process of converting computer programming language content into executable code cannot always be perfectly reversed, it is indisputably considered a translation or at least an adaptation of the former. The creation of such an adaptation or derivative work requires the authorization of the rights holder.[88] Lastly, it should be noted that the fact that some of the more advanced AI systems may be able to install "output filters" that may prevent outputs where large verbatim copies are generated, is of little consequence.[89] As explained above, copies consisting of numerical

---

v. Reno, Inc., 555 F.3d 11171, 1179 (10th Cir. 2009) ("Striking similarity exists when the proof of similarity in appearance is so striking that the possibilities of independent creation, coincidence and prior common source are, as a practical matter, precluded.") (internal citations and quotations omitted).

[85] *See supra* note I.

[86] *See supra* note I(C).

[87] Both source code and object code are treated as literary works under copyright law, for example *see* TRIPS Agreement, *supra* note 51, art 10.1,; Apple Comput., Inc. v. Franklin Comput. Corp., 714 F.2d 1240, 1249 (3d Cir. 1983), *cert. dismissed,* 464 U.S. 1033, 104 S. Ct. 690, 79 L. Ed. 2d 158 (1984) (source and object code); CMS Software Design Sys., Inc. v. Info Designs, Inc., 785 F.2d 1246, 1247 (5th Cir. 1986) (source code); Williams Elecs., Inc. v. Artic Int'l, Inc., 685 F.2d 870, 876-77 (3d Cir. 1982) (object code); Whelan Assoc. v. Jaslow Dental Laboratory, 797 F.2d 1222, 1233 (3d Cir. 1986) (object code and source code); and in Europe, Case C-406/10, SAS Institute Inc. v. World Programming Ltd, ECLI:EU:C:2012:259, (May 2, 2012) (object code and source code).

[88] *See* GOLDSTEIN, *supra* note 43, at 7.3.

[89] Output filtering is pertinent to several critical issues, including hate speech, misinformation, profanity, and privacy, among others. Copyright output filters, specifically, are designed to prevent the generation of content that infringes on copyrighted material.

representations of the training data are made and kept on the AI system regardless of whether the generation of infringing output is regulated at the point of exit.[90]

In conclusion, the inclusion of copyright-protected material in the training datasets of LLMs can result in the creation of unauthorized copies on two levels: temporary copies generated during the training process, and numerical representations of the training data embedded within the LLM after training. Both instances may lead to copyright liability, bearing in mind that the plaintiff need not show an intent to infringe on the defendant's part to win her case.[91]

2. Outputs

In addition to copyright liability for using copyrighted works as inputs without permission, there is a lot of discussion about how to treat outputs—those things generated by the AI systems built on training involving copyrighted works.

Understanding copyright liability for outputs generated by LLMs can be complex due to the multiple copyright rights involved. A prominent concern is the right of reproduction. The primary question here is whether the LLM has created something that is indistinguishable from, or substantially similar to, an existing copyrighted work. If so, infringement may occur unless an exception applies or the LLM did not have access to the original work.[92]

Another key right is the creation of derivative works, which includes adaptations or translations.[93] For example, consider an LLM that translates a recent Booker or Goncourt winning novel into another language, such as Japanese or Spanish.[94] This action would violate the right to translate, which is a specific aspect of the broader right to create derivative works.[95] It would also infringe the Berne Convention's exclusive translation right, because the Convention text "does

---

These filters typically rely on pattern recognition algorithms capable of detecting sequences of text that closely match known copyrighted works. For examples of content filtering, *see GenAI Content Filtering: How to Prevent Exposure of Sensitive Data*, NIGHTFALL FIREWALL FOR AI, https://docs.nightfall.ai/docs/content-filtering-sensitive-data-chatgpt.

[90] *See supra* Part I(B).

[91] *See* Buck v. Jewel-LaSalle Realty Co., 283 U.S. 191, 198 (1931) ("Intention to infringe is not essential under the [Copyright] Act.").

[92] For example, an unpublished paper manuscript. *See* 17 U.S.C. § 106(1) and 4 MELVILLE B. NIMMER & DAVID NIMMER, NIMMER ON COPYRIGHT § 13.03 ("[W]hat is required by the traditional standards of copyright law […] , for decades prior to adoption of the 1976 Act and unceasingly in the decades since, has included the requirement of substantial similarity.").

[93] *See supra* notes 66 and 67.

[94] The US Copyright Act provides an illustrative list of works that constitute a derivative work: translations, musical arrangements, dramatizations, fictionalizations, motion-picture versions, sound recordings, art reproductions, abridgments, and condensations. 17 U.S.C. § 101 (definition of "derivative work").

[95] *See id.*

not distinguish among means of translation."[96] In addition, such a translation could also violate the rights to reproduce and distribute the original work, not to mention potentially violate the moral rights of the author, especially if the translation uses material without proper attribution.[97] It is reasonable to expect that a court would not only enjoin the distribution of an unauthorized translation, but also potentially award damages to the rightful copyright holder. As discussed below, there could be a separate violation if rights management information is removed.[98]

This does not, however, fully answer hard questions about the right to prepare derivative works under US law. The Copyright Act provides an exclusive right "to prepare derivative works based upon the copyrighted work" and defines "derivative work" in part as any work "*based upon one or more preexisting works*." [99] This definition of the right could loosely be used as a *definition* of machine-learning when applied to the creation of literary and artistic productions. because AI machines can produce literary and artistic content (output) that is almost necessarily "based upon" a dataset consisting of preexisting works.[100] The definition cannot literally mean what it says because human creations are often, if not almost always, "based upon" some other work that the author has read, seen, consulted, experienced or been influenced by in some other way.[101] As Isaac Newton put it in a nutshell, we "stand on the shoulder of giants."[102]

The broad language of the first part of the statutory definition (the "based upon" clause) can be restrained by the enumeration that follows in application of the *ejusdem generis* rule.[103] One can argue that the list captures the major forms

---

[96] GINSBURG & RICKETSON, *supra* note 49, ¶11.27.

[97] *See* Jane C. Ginsburg, *The Most Moral of Rights: The Right to be Recognized as the Author of One's Work*, GEO. MASON. J. INT'L COMM. L. 44 (2016) (noting that a moral right of attribution on all categories of works is recognized in the copyright laws of Berne Convention member States other than the United States and that it is a U.S. obligation under Art. 6*bis* of the Berne Convention).

[98] 17 U.S.C. § 1202(a).

[99] 17 U.S.C. §§ 101,106(2) (emphasis added).

[100] Otherwise the LLM could not produce more content of this type, as explained in *supra* Part I(B).

[101] For example, it is well-known that to learn creative writing or art humans learn from existing masterpieces and other works. *See* Daniel Gervais, *The Derivative Right, or Why Copyright Protects Foxes Better than Hedgehogs,* 15 VAND. J. ENT. & TECH. L. 785, 851 (2013) ("By copying a master's work, the 'pupil' might at least get a glimpse of the great author's mind, which would seem like a normatively desirable process. '*L'art naît d'un regard sur l'art*,' as the French would say: art is born from a view on existing art.").

[102] *See generally* ROBERT K. MERTON, ON THE SHOULDERS OF GIANTS 8–12 (1993). Professor Bridy, for example, has argued along those lines "all cultural production is inherently derivative." Annemarie Bridy, *Coding Creativity: Copyright and the Artificially Intelligent Author*, 2012 STAN. TECH. L. REV. 5, 12 (2012).

[103] On the *ejusdem generis* rule, *see* Garcia v. United States, 469 U.S. 70, 74 (1984): When general terms follow an enumeration of persons or things, such general words are not to be

of derivation that come under the derivative work umbrella and that the opening clause may then just capture what has elsewhere been labelled "penumbral derivatives," which one could define as works covered by the broad opening words of the statutory definition of "derivative work" ("a work based upon one or more preexisting works") but not mentioned in the list of illustrations.[104] Other arguments to limit the reach of the right exist. This has been a long-standing question in copyright law. Professor Paul Goldstein, for example, has argued that, in light of the enumeration, the statutory text is intended primarily to protect certain licensing markets.[105] It can be argued that the massive copying of protected works to train and fine-tune LLMs constitutes a significant market for licensing, a matter to which the article returns below.

Another controversy that the production of literary and artistic material by LLMs elevates to a core issue is the originality controversy.[106] It is beyond cavil that, *to be protected* as a derivative work, a literary or artistic production must meet the originality condition applicable to other works of authorship but does that mean that, *to infringe* the derivative work right (belonging to a third party), the derivative work must also be original?[107] This has been a long-standing question in copyright. Professor Goldstein opined that the derivative work right may be infringed even if the derivative production would *not* qualify for protection as a work, and the Ninth Circuit agrees.[108] Professor Nimmer and the

---

construed in their widest extent, but are to be held as applying only to persons or things of the same general kind or class as those specifically mentioned. In 17 U.S.C. § 101 of course, the general words of the "based upon" clause precede instead of follow, but the canon could still be invoked. The canon, however, "cannot be used to 'obscure and defeat the intent and purpose of Congress' or 'render general words meaningless.'" United States v. Kaluza, 780 F.3d 647, 661 (5th Cir. 2015).

[104] For example mounting pages from art books on tiles as in Mirage Editions, Inc. v. Albuquerque A.R.T. Co., 856 F.2d 1341 (9th Cir. 1988).

[105] PAUL GOLDSTEIN, GOLDSTEIN ON COPYRIGHT § 7.3 (3d ed. 2012); Paul Goldstein, *Derivative Rights and Derivative Works in Copyright*, 30 J. COPYRIGHT SOC'Y U.S.A. 209, 221 (1983) (noting that "[i]t is no coincidence that the principal cases establishing broad rights against infringement by derivative works characteristically involve situations in which the alleged infringer had at some earlier point sought a license.").

[106] *See* Daniel Gervais, *AI Derivatives: The Application of the Derivative Work Right to Literary and Artistic Productions of AI Machines*, 52:4 SETON HALL L. REV. 1111 (2022) (discussing the requirement applied by some US courts that a defendant's production be original to qualify as an infringing derivative work). Interestingly, even if the defendant's production is original, it would generally not be protected by copyright if it is infringing under 17 U.S.C. § 103(a).

[107] *See id.*

[108] Goldstein, *supra* note 105, at 231 n.75 (1983) ("[T]he Act does not require that the derivative work be protectable for its preparation to infringe."); *see also* Mirage Editions, Inc. v. Albuquerque A.R.T. Co., 856 F.2d 1341, 1342 (9th Cir. 1988); Munoz v. Albuquerque A.R.T. Co., 38 F.3d 1218 (9th Cir. 1994). In a 1909 Act case, the Ninth Circuit found that it made "no difference that the derivation may not satisfy certain requirements for statutory copyright registration itself." Lone Ranger Television, Inc. v. Program Radio Corp., 740 F.2d 718, 722 (9th Cir. 1984).

Seventh Circuit have taken a different view, however, though not in the context of AI.[109]

Legal adaption of the copyright framework to LLMs will happen in several ways. An amendment to the copyright statute is only one of them. [110] Courts will also play their customary role. [111] As of this writing more than thirty court cases were pending internationally to determine in particular the scope of exceptions such as fair use in the United States or the 2019 EU Directive.[112] Then, private ordering is likely to play a prominent role to put an end to or avoid litigation and increase certainty for all parties involved, probably taking the form of licensing arrangements that would determine what can, and cannot be done with copyrighted material used for commercial training purposes. The production of certain derivative works may be a prime target for such a contractual vehicle, considering the fuzziness of the borders of the derivative work right.

A final issue with a number of LLMs is that they are trained on large amounts of data, such as material available online.[113] In this case, even if the creators of the LLM claim they may not know exactly what the model was trained on, but it can be argued that they knew or should have known that some portion of the material was copyrighted.[114] Then, whether rights in a particular work in the dataset have been infringed will presumably follow the traditional infringement analysis.[115]

3. Rights Management Information

The WCT and the WPPT added an obligation to provide "adequate and effective remedies against any person who knowingly" removes or alters "electronic rights management information" without authority, or distributes, imports for distribution, broadcasts or communicates to the public works or copies of works without authorization, knowing that electronic rights management

---

[109] Lee v. A.R.T. Co., 125 F.3d 580, 582 (7th Cir. 1997); MELVILLE B. NIMMER & DAVID NIMMER, 1 NIMMER ON COPYRIGHTS §§ 3.01-3.03.

[110] CHRISTOPHER T. ZIRPOLI, CONG. RESEARCH SERV., LSB10922, GENERATIVE ARTIFICIAL INTELLIGENCE AND COPY. LAW (Sept. 30, 2023) ("Congress may consider whether any of the copyright law questions raised by generative AI programs require amendments to the Copyright Act or other legislation.").

[111] *See id.* ("Given how little opportunity the courts and Copyright Office have had to address these issues, Congress may adopt a wait-and-see approach. As the courts gain experience handling cases involving generative AI, they may be able to provide greater guidance and predictability in this area through judicial opinions.").

[112] *See infra,* Parts III. A and B.

[113] *See supra* Part I.B.

[114] *See* Sean Hollister, *Microsoft's AI boss thinks It's Perfectly Okay To Steal Content If It's On The Open Web*, THE VERGE (June 28, 2024), www.theverge.com/2024/6/28/24188391/microsoft-ai-suleyman-social-contract-freeware.

[115] It is not necessary to show that the defendant intended to copy a specific work. The case law has recognized "unconscious crying" as sufficient, for example. *See e.g.*, ABKCO Music, Inc. v. Harrisongs Music, Ltd., 722 F.2d 988, 998 (2d Cir. 1983).

information has been removed or altered without authority.[116] "Rights management information" means "information which identifies the work, the author of the work, the owner of any right in the work, or information about the terms and conditions of use of the work, and any numbers or codes that represent such information, when any of these items of information is attached to a copy of a work..."[117]

This means that the removal of rights management information, which often occurs during the training of a model, could constitute a separate violation of a copyright holder's rights[118]. However, to prove infringement, the plaintiff must show that the defendant knew or had reasonable grounds to know that the removal would "induce, enable, facilitate or conceal an infringement" of a copyright, such as reproduction. There is very little case law on this type of infringement, but it is likely to be considered by the courts in a number of pending cases in the United States (where the WCT obligations in this respect have been implemented in Chapter 12 of Title 17).[119]

III. LLM-RELATED EXCEPTIONS AND LIMITATIONS IN NATIONAL LAWS

Part I above contends that incorporating copyright-protected material in the training datasets of LLMs may result in the creation of unauthorized copies, potentially leading to prima facie copyright liability. Therefore, it is crucial to examine whether such liability may be mitigated through the application of relevant copyright exceptions.

Exceptions and limitations, like the exclusive rights themselves, are carefully crafted to promote the overall copyright system.[120] There is no single set of global exceptions and limitations. Instead, each country fashions its own specific conditions. All countries that are parties to the Berne Convention (TRIPS) (that is to say, almost all countries), however, must abide by the Convention's "three-step test."[121] This test was initially designed to guide the implementation of exceptions

---

[116] WCT, *supra* note 52, art.12; WPPT *supra* note 52, art.19.

[117] *Id.*

[118] Typically, the metadata is simply not copied, so that the metadata is present in the original copy but not in the training copy. Some lawsuits have argued that this is not removal, but rather "not copying". But the fact that a copy is made that differs from the original in that the metadata is no longer present makes the argument seem like a distinction without a difference

[119] *See* ADR Int'l Ltd. v. Inst. for Supply Mgmt Inc., 667 F. Supp. 3d 411 (S.D. Tex. 2023); O'Neal v. Sideshow, Inc., 583 F. Supp. 3d 1282 (C.D. Cal. 2022); Sid Avery & Assocs., Inc. v. Pixels.com, LLC, 479 F. Supp. 3d 859 (C.D. Cal. 2020); Splunk Inc. v. Cribl, Inc., 662 F. Supp. 3d 1029 (N.D. Cal. 2023); Logan v. Meta Platforms, Inc., 636 F. Supp. 3d 1052 (N.D. Cal. 2022).

[120] GERVAIS, *supra* note 40, at 216-30 (discussion on role of exceptions).

[121] Berne Convention supra note 51, art. 9(2); TRIPS Agreement art. 13; for a discussion on the application of the three step test in national laws *see* Christophe Geiger et al., *The Three-Step-Test Revisited: How to Use the Test's Flexibility in National Copyright Law*, 29 AM. UNIV. INT'L L. REV. 581 (2014).

to the reproduction right, and TRIPS further extended it to other exclusive rights under copyright.[122] Given that the United States, European Union, United Kingdom, Japan, Singapore and Switzerland are all World Trade Organization (WTO) members, the test is relevant across these jurisdictions. Specifically, the test mandates that members limit exceptions to "certain special cases" that do not interfere with the normal use of the work nor unjustly harm the legitimate interests of the rights holder.[123]

While the test may be considered ambiguous, it has been construed as follows: (1) exceptions may not be overly broad, thus applicable only in "certain special cases"; (2) exceptions may not "rob right holders of a real or potential source of income that is substantive" which would "conflict with normal exploitation of the work"; and (3) exceptions may not "do disproportional harm to the rights holders," hence may not "prejudice legitimate interests" of the right holder.[124] It is useful to have the three step test in mind when considering the potential scope of national exceptions in the context of AI training.[125]

A. United States

Many proponents of AI have contended that the U.S. concept of fair use, codified in 17 U.S.C. 107, covers many uses of copyrighted materials in the LLM process. The nature of fair use arguments advocated by AI companies in relation to training appear to focus particularly on the first factor: "*the purpose and character of the use, including whether such use is of a commercial nature or is for non-profit educational purposes.*"[126] In particular such arguments seek to create parallels between earlier cases such as *Google Books*,[127] which are said to involve "machine learning" and present disputes involving AI training. Under the first factor, AI companies appear to rely on their claim that their use of copyrighted material in training their AI models is "transformative."[128]

---

[122] DANIEL GERVAIS, TRIPS AGREEMENT: DRAFTING HISTORY AND ANALYSIS (5th ed. 2021).

[123] P. Bernt Hugenholtz & Ruth Okediji, *Conceiving an International Instrument on Limitations and Exceptions to Copyright* (Amsterdam L. Sch. Rsch. Paper No. 2012-43, Inst. for Info. L. Rsch. Paper No. 2012-37), https://ssrn.com/abstract=2017629.

[124] *Id.*

[125] As the three-step test is derived from international treaties, as such it does not directly apply to national law. However, courts may use it as an interpretative tool when construing exceptions in national law. Whether they do so is subject to debate in each instance. For example, *see* Barry Sookman *The Google Book Project: Is It Fair Use?*, 61 J. COPYRIGHT SOC'Y 485, 485-515 (2014) (critiquing the court decision in the Authors Guild v. Google Inc., 770 F. Supp. 2d 666 (S.D.N.Y. 2011), and arguing that the decision may not be compatible with the three-step test).

[126] *See, e.g., OpenAI and journalism*, OPENAI.COM (Jan. 8, 2024), https://openai.com/index/openai-and-journalism/ (commenting on the NYT's lawsuit, where OpenAI seeks to emphasize the "transformative potential" of AI).

[127] Authors Guild v. Google, Inc., 804 F.3d 202 (2nd Cir. 2015).

[128] *See*, *e.g.*, OpenAI, LP, Comment Regarding Request for Comments on Intellectual Property Protection for Artificial Intelligence Innovation (Dec.16, 2019), https://www.uspto.gov/sites/default/files/documents/OpenAI_RFC-84-FR-58141.pdf.

The debate on whether using copyrighted works to train AI systems constitutes fair use, particularly under the transformative use doctrine, is both complex and novel. In *Campbell v. Acuff Rose Music*, for example, the Supreme Court highlighted that the distinction between mere duplication and transformative use is crucial.[129] Duplication likely causes market harm, whereas transformative uses may not directly compete with the original and could be considered fair use.[130] The term "transformative" suggests a significant alteration or a new purpose different from the original, raising questions about AI training's qualification under this criterion. In *Sony*, the Court applied the fair use doctrine to a new technology at the time (the VCR) with "substantial noninfringing use" that would transform the movie industry.[131]

AI companies posit that their AI training process is highly transformative and can be used for noninfinging purposes, arguing in addition that, while original works are created for human consumption, their use in AI training is for non-expressive purposes—to enable AI systems to learn human-generated patterns, serving a different objective and resulting in outputs not intended for direct human consumption.[132] Those claims, however, invite scrutiny. Generative AI systems, comprise two distinct software layers, the input layer or encoder - ingesting existing works and the output layer or decoder - producing results. It is a process that, despite being for training purposes, still utilizes the copyrighted material for its inherent expressive value when ingesting it into the first "layer." After "harvesting" its inherent expressive value, the system may discard the copy when it no longer needs it. As this Article sees it, the copying occurring when the work is ingested into the system's first "layer" is taking place exactly due to the work's expressive value, which parallels the underlying reason for human consumption.

The ingestion of works into generative AI's encoder layer closely corresponds to how humans learn from copyrighted materials in professional workbooks. They may wish to consult such books that contain copyrighted works in order to learn new styles and techniques that are to be found therein, but they are not likely to be excused under copyright law if they choose to make unauthorized copies of such books in the process of such learning. Use that involves copying of books, even if for learning purposes, is likely to require a license.

---

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

[130] *See id.* The idea of competition with the original was central to the Supreme Court's analysis sunder the first factor in Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 598 U.S. 508, 556 (2023) ("[U]nder the first fair-use factor the salient point is that the purpose and character of the Foundation's use involved competition with Ms. Goldsmith's image.") (Gorsuch J, concurring).

[131] *Sony*, *supra* note 57. Prior to the VCR, the primary revenue stream for the movie industry was from theatrical box office sales. However, the introduction of the VCR allowed consumers to purchase or rent movies for home viewing, creating a new revenue stream for the industry. This shift led to the development of the home video market, which became a lucrative source of income for movie studios.

[132] OpenAI, LP, *supra* note 128, at 4-5.

The Court of Appeals for the Second Circuit's decision in *American Geophysical Union v. Texaco*, *Inc.* further challenges AI companies' position in the present context, demonstrating that copying for research or educational purposes without transformation is not likely to constitute fair use.[133] Similarly, search engine and indexing cases like Kelly v. Arriba Soft Corp., HathiTrust, and Authors Guild v. Google, where the contested uses were considered transformative and hence fair use, differ significantly from AI training.[134] In those cases, the use served a new purpose—facilitating information access without replacing the original works' consumption. The fact that the works were not used for their expressive or aesthetic value appears to have weighed heavily in favor of fair use. As mentioned, however, in the case of AI training, the ingestion into the encoder "layer" of the system is occurring precisely due to its aesthetic or expressive value.

At the heart of the first fair use factor is the assessment of whether an AI's system's ingestion of copyrighted works serves a new, valuable purpose or merely repurposes the original's expressive content. As Part II explains, when discussing "tokenization" and "word embeddings," training AI models involves a deep engagement with the creative expressions of copyrighted works, aiming to internalize and replicate their stylistic and compositional principles for generative purposes. This direct engagement with the copyrighted expression for training contrasts with search engines' incidental use of copyrighted material to facilitate access to original works.

Finally, it should be noted that the reliance of AI companies on the contested use being transformative appears less tenable in light of the recent Supreme Court decision in *Andy Warhol Foundation for the Visual Arts, Inc. v Goldsmith*.[135] In that case, the Supreme Court suggested that the key determination under the first fair use's first factor relates to the nature of the challenged use. The fact that such use may be commercial in nature is also important; while not a decisive issue, commercial use weighs against fair use. Importantly, according to Warhol when it comes to the nature of the challenged use, it is necessary to determine its purpose. If the challenged use has a similar purpose to that of the original, the entire first factor is likely to weigh against fair use. As mentioned above, the reason for ingesting copyrighted works into the encoder layer of an AI model is directly and unequivocally due to its expressive and aesthetic value. It is due to their expressive and aesthetic value that such works are put into the marketplace. Hence, the decision in *Warhol* further complicates the position of AI companies under the first fair use factor.

---

[133] 60 F.3d 913 (2d Cir. 1994). AI companies frequently present a case similar to Texaco's unsuccessful stance, asserting that the training of AI systems undeniably promotes scientific investigation, thereby contributing to the progress of arts and sciences. *See* Am. Geophysical Union v. Texaco Inc., 60 F.3d 913 (2d Cir. 1994).

[134] Kelly v. Arriba Soft Corp., 336 F.3d 811 (9th Cir. 2003); Authors Guild, Inc. v. Hathitrust, 755 F.3d 87, 94 (2d Cir. 2014); Authors Guild v. Google, Inc., 804 F.3d 202 (2nd Cir. 2015).

[135] 598 U.S. 508 (2023).

This position does not necessarily improve upon assessing the three remaining fair use factors. The second and third factors, the nature of the copyrighted work and the amount and substantiality of the portion used, appear to weigh clearly against fair use. The works in question are mostly highly expressive works and are copied and ingested in full. Lastly "the effect of the use on the potential market for or value of the work," could also weigh against fair use. This is particularly so in light of the availability of licenses and AI companies' growing tendencies to enter licensing arrangements with copyright holders.[136] For example, it has been widely reported that OpenAI entered a license with Axel Springer, News Corp, and Associated Press to use their articles in its products.[137] Such practices make any argument as to lack of impact on the plaintiff's market less convincing. If there is a market for licensing copyrighted work for ingestion in AI models, as clearly there is, then using such works in that way and without authorization clearly deprives the rightsholder of potential licensing fees. Notably, it was reported that OpenAI and the New York Times engaged in discussions to reach a commercial agreement for the use of the New York Times content. However, unlike the cases mentioned earlier, these negotiations were unsuccessful. As a result, the New York Times sued OpenAI and Microsoft for copyright infringement.[138]

In sum, while we are probably years away from final rulings in some of the disputes that are presently pending, AI companies' position under the fair use doctrine appears to be precarious, taking into account fair use jurisprudence and AI companies' technological and business models.

B. European Union

The position in the EU is somewhat more straightforward than that of the United States, as copyright exceptions in instances involving text and data mining (TDM) that seem applicable to the training of AI models are explicitly provided for under EU legislation.

---

[136] Michael M. Grynbaum & Ryan Mac, *The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work*, NEW YORK TIMES (Dec. 27, 2023), https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html.

[137] Daniel Thomas & Madhumita Murgia, *Axel Springer strikes landmark deal with OpenAI over access to news titles*, FINANCIAL TIMES (Dec. 13, 2023), https://www.ft.com/content/7cd439bc-29cd-44f9-8676-4761e27bc3a8; Matt O'Brien, *ChatGPT-maker OpenAI signs deal with AP to license news stories*, ASSOCIATED PRESS (Jul. 13, 2023, 8:41 AM), https://apnews.com/article/openai-chatgpt-associated-press-ap-f86f84c5bcc2f3b98074b38521f5f75a; *News Corp and OpenAI Sign Landmark Multi-Year Global Partnership*, NEWS CORP (May 22, 2024), https://newscorp.com/2024/05/22/news-corp-and-openai-sign-landmark-multi-year-global-partnership/.

[138] *Supra* note 136.

1. Digital Single Market 2019

TDM involving copyrighted works is largely governed under the EU Directive on copyright and related rights in the Digital Single Market 2019 (DSM).[139] Article 3 of the Directive introduces a new copyright exception for "reproductions and extractions conducted by research organizations and cultural heritage institutions for the purpose of conducting scientific research through text and data mining of works or other subject matter to which they lawfully have access."[140] Additionally, Article 2 of the Directive clarifies the definition of a research organization as a university, its libraries, or any research entity primarily aimed at conducting scientific research or educational activities that include scientific research, on a non-commercial basis (including reinvesting all profits back into scientific research) or in pursuit of a public interest mission recognized by a member state.[141] Thus, as long as the organization at issue is non-commercial, the substantive requirements to be complied with are having lawful access to the works and keeping any reproductions or extractions for no more than the duration necessary to achieve the objectives of the TDM being conducted.[142]

Article 4 extends this exception to commercial entities for the purpose of TDM and adds that, in addition to the requirements of Article 3, it must also be the case that the right holders have not opted out by making the appropriate reservation to that effect.[143] If this is the case, commercial companies may engage in TDM while sheltering under the exception in Article 4[144]. As the Directive has

---

[139] Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC [hereinafter DSM Directive]. It is noteworthy that this EU legislation is distinct from the AI Act and was enacted five years earlier. The AI Act makes several references to the DSM Directive in various instances.

[140] *Id.*, art. 3.

[141] *Id.,* art. 2.

[142] It is noteworthy that the German implementation of the DSM Directive provides that research organizations risk losing their exemption under Article 3 of the if they partner with private enterprises that wield influence or receive preferential access to their research findings (S. 60(d)(2) UrhG). This measure is intended to ensure that TDM exemptions are not misused for commercial gain.

[143] DSM Directive, *supra* note 139, art. 4. This seems to mean that copies created as a result of TDM activities should not be retained longer than necessary for the TDM process. For example, any temporary copy generated during the process must remain so, and not be kept as a permanent training dataset.; While the Directive mentions machine readable reservation, it is not yet entirely clear how this may work in practice.

[144] It is noteworthy that a non-profit research organization that may benefit from Art. 3, is defined under Art.2 of the DSM Directive as follows: "'research organization' means a university, including its libraries, a research institute or any other entity, the primary goal of which is to conduct scientific research or to carry out educational activities involving also the conduct of scientific research:

(a) on a not-for-profit basis or by reinvesting all the profits in its scientific research;

(b) or (b) pursuant to a public interest mission recognized by a Member State;

only been in force for a relatively short period of time, it is not yet clear how these provisions will be interpreted by the courts, and to what extent activities related to the training, fine tuning, and use of LLMs constitute a form of TDM. Similarly, it remains uncertain how a right holder could feasibly and without undue burden opt-out.

2. AI Act

The EU AI Act aims to form part of a comprehensive EU legal framework on AI, addressing the associated risks. It provides AI developers and deployers with clear requirements and obligations for specific AI uses. This Act is intended to be part of a broader policy initiative to promote trustworthy AI, ensuring the safety and fundamental rights of individuals and businesses in relation to AI, and enhancing uptake, investment, and innovation in AI across the EU.

While most of the issues governed under the AI Act have little to do with copyright liability or intellectual property law, the Act does introduce a number of changes that are relevant here. The Act provides for a disclosure obligation, where providers of foundations models have to draw up and make publicly available a sufficiently detailed summary of the content used for training of the foundation model, according to a template provided by the AI Office.[145] Thus, the Act provides, in principle, for a mechanism that would enable right holders to ascertain whether their works have been used for training specific AI models. As the AI Act passed as recently in March 2024, it is yet to be seen how exactly this disclosure requirement will come into effect. Finally, the EU has taken steps to discourage the creation of AI training safe heavens, designed to circumvent the aforementioned EU rules on TDM provided for under the DSM Directive. Recital 106 of the AI Act stipulates that providers of AI models should put in place a policy to respect EU law on copyright and related rights, in particular to identify and respect any opt-outs expressed by right holders pursuant to Article 4(3) of DSM Directive, regardless of the jurisdiction in which the copyright-relevant acts underpinning the training of these AI models take place.[146] Namely, as long as an AI model is offered in the EU, it must comply with the DSM Directive's Articles 3 and 4, regardless of the jurisdiction in which it was trained. Reiterating the aforementioned position under the DSM Directive, the AI Act stipulates that in relation to general purpose AI models, any use of copyright-protected content necessitates the authorization of the respective rights holders, unless a suitable

---

in such a way that the access to the results generated by such scientific research cannot be enjoyed on a preferential basis by an undertaking that exercises a decisive influence upon such organization"; Any organization that does meet this definition must comply with Art. 4.

[145] Proposal for a Regulation of the European Parliament and of Council Laying Down Harmonized Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, COM (2021) 206 final (Mar. 13, 2024) [hereinafter AI Act].

[146] *Id.*

copyright exception applies.[147] Pertinent to the DSM Directive, it specifies that rights holders can reserve their rights over their copyright works or other subject matter to prevent TDM, except when conducted for noncommercial scientific research purposes. If rights have been explicitly reserved via an appropriate machine-readable "opt out," providers of foundational models must obtain authorization from the rights holders to perform TDM on such works.

Significantly, consideration should also be given to the model as trained. The EU's AI Act sheds light on the applicability of the DSM Directive's TDM provisions to AI training, confirming that these provisions permit reproductions and extractions of works or other subject matter TDM purposes under specified conditions.[148] However, the exemptions provided under the TDM provisions apply primarily to the right of reproduction and do not extend to the right of communication to the public, which includes the right of making available. Consequently, developers of AI models who comply with Articles 3 and 4 of the DSM Directive may be exempt from copyright liability for reproductions related to temporary copies created during the training process and permanent copies embedded within the models. However, this exemption does not necessarily extend to other copyright restricted acts that fall outside the scope of TDM exemptions. The central question regarding the provision of AI modules to the public is whether such provision constitutes "making available" within the context of the communication to the public right under Article 3 of the Information Society Directive. If this is the case, Articles 3 and 4 of the DSM Directive may not apply.[149] Article 3(1) of the Information Society Directive states: "Member States shall provide authors with the exclusive right to authorize or prohibit any communication to the public of their works, including the making available to the public of their works in such a way that members of the public may access them from a place and at a time individually chosen by them." This provision aims to ensure that authors retain control, inter alia, over the accessibility of their works, particularly in the digital environment.

3. The AI Act and Making Available Right

The key issue is therefore the public availability of, and accessibility to, the unauthorized copies of works used as training material, or fragments thereof, contained within AI models once trained. This presents a novel question in the context of the EU communication to the public right, which has yet to be addressed by the courts. Two cumulative criteria must be present for the right of communication to the public to be engaged. First, the act of "communication" must take place, and secondly, the communication of a work must be to a "public." In the present context, the second criterion does not pose a unique challenge as members of the public to which an AI models, such as ChatGPT, are available constitute a "public" for the purposes of Article 3. It is the first criterion that

[147] *Id.*, Art. 53(1c).

[148] *Id.*

[149] *See id.*

requires closer scrutiny. The jurisprudence of the Court of Justice of the European Union (CJEU) provides that it is sufficient for a "communication" to take place where a work is made available to the public in such a manner that members of that public can access it, regardless of whether they actually do so. Thus, it is necessary to establish whether having an AI model available for public use, for example - in a manner akin to subscription services offered by AI companies such as ChatGPT, constitutes public accessibility to unauthorized copies contained within the model, within the meaning of the communication to the public right. An analysis of CJEU decisions such as *Filmspeler* and *Pirate Bay* suggests that the answer may be affirmative, at least where what is made available is one of more preexisting works.

In the *Filmspeler* case[150] the CJEU found that offering a media player pre-loaded with add-ons that provided links to illegal streams of copyrighted content amounted to making the works available to the public, as it enabled users to access the content without the right holder's permission. Similarly, in the *Pirate Bay* case[151] the CJEU held that the operators of The Pirate Bay website facilitated access to protected works, making them available to the public by indexing and categorizing torrent files; this was sufficient to constitute a communication to the public within the meaning of Article 3 of the Information Society Directive. In both cases, offering the public a vehicle through which they may access unauthorized copies, or fragments thereof, amounted to making such copies available to the public within the meaning of the communication to the public right under Article 3 of the Information Society Directive. We have observed that AI models are "memorizing" or effectively storing portions of copyrighted works used as training material. This allows them, under certain circumstances, to generate outputs that include such portions. Following the rationale adopted by the CJEU, the ability for members of the public to access these models, prompt them, and trigger the generation of such outputs may be considered a facilitation of public access to infringing copies, similar to the situations in *Filmspeler* and *Pirate Bay*. Thus, enabling public access to AI models that can reproduce portions of copyrighted works can be argued to constitute making infringing copies available to the public when the outputs contain substantial portions of preexisting works, in line with the CJEU's interpretation in these cases.

4. The Potential Long-Reach of the AI Act

Like other areas of intellectual property law, copyright law operates on the principle of territoriality. This means the location of an infringement typically determines the applicable law. Consequently, if an AI model is trained in a specific country, the permissibility of such training is governed by that country's copyright laws. This situation could potentially lead to forum shopping, where AI companies

[150] Case C-527/15, Stichting Brein v. Jack Frederik Wullems, ECLI:EU:C:2017:300, (Apr. 26, 2017).
[151] Case C-610/15, Stichting Brein v. Ziggo BV, ECLI:EU:C:2017:456, (June 14, 2017).

train their models in jurisdictions with more lenient copyright rules and then offer these models in countries with stricter regulations.

The AI Act addresses this potential strategy to ensure a "level playing field." It introduces provisions to counter forum shopping in Recital 106.[152] Article 53(1)(c) of the AI Act then further confirms that European Union law on copyright and related rights, as referenced in Recital 106, includes the TDM exceptions of the DSM Directive, particularly Article 4(3). These provisions have the potential to create a "Brussels effect" similar to that of the EU General Data Protection (GDPR) rules.[153] Essentially, EU copyright law regarding the training and operation of AI models would need to be followed globally, as long as an AI model is marketed in the EU. Given that the EU represents the largest economic market in the world, many AI companies may find it most economically viable and practical to design and train their AI models in compliance with EU copyright laws.[154]

C. United Kingdom

Current UK law essentially mirrors Article 3 of the EU DSM Directive. Under Section 29A CDPA1988, the creation of copies done "for the sole purpose of research for a non-commercial purpose" is excused under copyright law.[155] There is no equivalent of Article 4; copies created by commercial entities and for commercial purpose would likely constitute copyright infringement, as none of the general UK copyright exceptions is likely to apply to AI training by commercial entities. Hence, copying resulting from AI training for commercial purposes requires a license. The discussion of the applicability of the EU's TDM exceptions to reproductions arising from the creation of permanent copies

---

[152] Recital 106 of the AI Act provides: "Providers that place general-purpose AI models on the Union market should ensure compliance with the relevant obligations in this Regulation. To that end, providers of general-purpose AI models should put in place a policy to comply with Union law on copyright and related rights, in particular to identify and comply with the reservations of rights expressed by rightsholders pursuant to Article 4(3) of Directive (EU) 2019/790. Any provider placing a general-purpose AI model on the Union market should comply with this obligation, regardless of the jurisdiction in which the copyright-relevant acts underpinning the training of those general-purpose AI models take place. This is 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."

[153] The term "Brussels Effect" is derived from the "California Effect" of California emission standards for cars that became de facto standards throughout the United States. This is arguably what happens with EU rules concerning GDPR's protection of personal data. *See* Joanne Scott, *The Global Reach of EU Law*, *in* EU LAW BEYOND EU BORDERS (Marise Cremona & Joanne Scott, eds., 2019).

[154] On the EU being the largest economy in the world, *see The EU Position in World Trade*, EUROPEAN COMM'N, https://policy.trade.ec.europa.eu/eu-trade-relationships-country-and-region/eu-position-world-trade_en.

[155] Copyright, Designs and Patents Act 1988, c. 47, § 29A (UK) [hereinafter CDPA].

embedded in the "trained" model can be similarly applied to Section 29A of the CDPA. It is likely that such reproductions are excused under the conditions set forth in Section 29A. However, like the TDM provisions, Section 29A does not seem to cover infringements that arise from violations of the making available right under Section 20(2)(b) of the CDPA.[156]

The UK does not have a jurisdiction shopping busting provision similar to Recital 106 of the EU AI Act. Consequently, where the copyright-relevant acts take place outside the UK, copyright liability is not likely to be triggered. Nevertheless, since we have seen that it is not only the act of training that is likely to attract copyright liability, but also the model as trained, it is possible that the mere offering of a trained model, which contains unauthorized copies or adaptations of copyrighted works, may constitute copyright infringement irrespective of the territory in which the actual training took place.[157]

At present, there is a major lawsuit before the UK courts concerning copyright infringement liability of an AI foundation model developer and provider.[158] The court is examining issues related to the liability for using copyright-protected material during the training and development of an AI model, as well as potential infringement resulting from the outputs generated by such models.[159]

D. Japan

The Copyright Law of Japan was amended in 2009—that is, before the emergence of LLMs—to allow computerized "information analysis" defined as "to extract information, concerned with languages, sounds, images or other elements constituting such information."[160] The exception distinguishes the

---

[156] *Id.* §20(2)(b).

[157] For example, the aforementioned 'word embeddings' are likely to constitute unauthorized adaptations, as they 'capture' the essence of copyrighted works' expressive and aesthetic value.

[158] Getty Images (US) Inc & Ors v Stability AI Ltd, [2023] EWHC 3090 (Ch).

[159] In legal proceedings before the High Court, Getty Images has accused Stability AI of infringing its intellectual property rights by using its images without authorization to train the AI model, Stable Diffusion. Getty Images also claims the outputs from Stable Diffusion reproduce significant parts of its copyrighted works or bear Getty's brand. Additionally, Getty Images alleges secondary copyright infringement, arguing that Stable Diffusion constitutes an "article" under the Copyright, Designs and Patents Act 1988 (CDPA) sections 22, 23, and 27, which Stability AI knew or had reason to believe was an infringing copy of the work. Stability AI's application to dismiss these claims was rejected, and the case will proceed to trial in summer 2025. Stability AI's defense includes arguments that the training occurred outside the UK and thus fall outside the scope of UK copyright law and that its outputs are created without reproducing specific images from the training data and that any act of copying is solely the user's responsibility. Furthermore, regarding the outputs it also sought to invoke the pastiche exception under Section 30A CDPA.

[160] *Copyright Law of Japan*, COPYRIGHT RSCH. AND INFO. CTR. (CRIC) (Jan.19, 2023), https://www.cric.or.jp/english/clj/index.html.

analysis of information from the "purpose of enjoying the thoughts and/or feelings expressed in the copyrighted work."[161]

Recent guidance issued by the government agency responsible for administering copyright law indicates that the consent of the copyright holder is not required unless there is a "material impact on the relevant markets" and the use of AI does not "infringe the interests of copyright holders."[162] This guidance is consistent with the application of the three-step test.[163] Thus, training a model is prohibited "if the intention is to output products that can be perceived as creative expressions of copyrighted works, including imitating the style of certain creators."[164]

E. Singapore

In Singapore, the Copyright Law was amended in 2021 to make an exception to the rights of reproduction and communication to allow "computational data analysis" (CDA).[165] CDA includes "using a computer program to identify, extract and analyze information or data from the work or recording," and "using the work or recording as an example of a type of information or data to improve the functioning of a computer program in relation to that type of information or data." There are limits to this exception In particular, the user must have lawful access and the copy must be "made for the purpose of CDA or "preparing the work or recording for" CDA; and, importantly, the user must not "use the copy for any other purpose."[166] There are also permitted forms of communication to the public, with specific restrictions. It is worth noting that any "contract term is void to the extent that it purports, directly or indirectly, to exclude or restrict" the uses permitted by this exception.[167]

F. Switzerland

In 2019, Switzerland amended in federal copyright statute to allow "[f]or the purposes of scientific research, [the reproduction pf] a work if the copying is due to the use of a technical process and if the works to be copied can be lawfully accessed."[168]

---

[161] *Id.*

[162] *Id.*

[163] *Id.*

[164] Scott Warren & Joseph Grasser, *Japan's New Draft Guidelines on AI and Copyright: Is It Really OK to Train AI Using Pirated Materials?*, PRIVACY WORLD (Mar. 12, 2024), https://www.privacyworld.blog/2024/03/japans-new-draft-guidelines-on-ai-and-copyright-is-it-really-ok-to-train-ai-using-pirated-materials/.

[165] Copyright Act 2021 (Act No. 22 of 2021, amended by the Statutes (Miscellaneous Amendments) Act 2022), §§ 243-244.

[166] *Id.* § 244(2).

[167] *Id.* § 187.

[168] BUNDESGESETZ ÜBER DAS URHEBERRECHT UND VERWANDTE SCHUTZRECHTE [FEDERAL ACT ON COPYRIGHT AND RELATED RIGHTS], Oct. 9, 1992, SR 231.1 (Switz.).

IV. LICENSING AS A KEY PART OF THE PATH FORWARD

There is considerable uncertainty about the future. This is potentially costly both to AI companies, which may end up paying large statutory and compensatory damages in multiple jurisdictions, and to authors and other rights holders whose livelihoods are directly affected by LLMs. The fair use debate in the United States is likely to continue for several years until one or more Supreme Court opinions shed additional light on the issue. While LLMs are a significant and new technology and may be capable of multiple non-infringing uses, not every use of them with copyrighted material is transformative. AI companies know this. For example, an LLM that ingests scientific articles and then simply regurgitates them on demand would not perform a transformative function. Nor would ingesting, say, all the music written by Taylor Swift for the purpose of producing more (but free) music "like her" in a way that infringes on her rights in the ingested musical works be transformative in our view, especially in light of the Warhol decision. Nevertheless, some uses of LLMs and their training may be found to be fair use.

Even in the EU, where more specific legislation has been adopted, debate continues about the scope of the TDM exceptions, their interplay, their interface with the EU AI Act, and the operation of the Article 4 opt-out. In other jurisdictions, limitations on statutory exceptions, such as in Japan and Singapore, may also require additional clarification.

The rapid change and uncertainty in the realm of AI and copyright raises the inevitable question of how to legally enable users to access high quality and compliant materials for use in AI systems, given the variability and attendant uncertainty about the scope of rights and exceptions and limitations.[169] Where does the law come down on the creation of LLMs, both in the input and output of existing copyrighted materials? The answer to this conundrum may simply lie in the time-tested solution that has proven successful during earlier periods of technological advancement: licensing.[170] Licensing enables copyright owners and users to come together in a mutually beneficial manner, helping the market function more efficiently and responsibly.

There is no single global copyright law, and countries vary significantly in their approach to copyright and AI-related issues like text and data mining and transparency.[171] There is also no single court that will hand down all decisions on copyright and AI, either within a country or globally as the text of exceptions and limitations in national law varies greatly However, global licenses can harmonize how copyright owners and users agree to use copyrighted works, significantly benefiting innovation and progress by setting the stage for consistent and responsible copyright uses that could lead to untold scientific and cultural advancements. Licenses could put an end to much of the uncertainty and to both

---

[169] *See supra* Part III.

[170] *See* DANIEL GERVAISE, RESTRUCTURING COPYRIGHT: A PATH TOWARDS INTERNATIONAL COPYRIGHT REFORM 231-56 (Revised and updated ed. 2019).

[171] *See supra* Part III.

pending and potential future litigation, putting acceptable boundaries on what can and cannot be done with copyrighted material when training LLMs.

Various licensing models could play a crucial role in this progress. Direct licensing—agreements between a copyright owner and user—is incredibly important because it allows the parties to be flexible in defining terms like payment, timing, and addressing specific, bespoke use cases.[172] Voluntary collective licensing is also likely to play a critical role in solving the licensing puzzle, enabling users to obtain a single license that can cover thousands (or more) copyrighted works without having to negotiate with each copyright owner individually.[173] This approach is highly beneficial for both copyright owners and users, as it provides an efficient mechanism to grant and obtain permission for using copyrighted works.[174]

Voluntary collective licensing is uniquely equipped to handle some of the more complex issues, when there are large numbers of works and potential users searching for an efficient mechanism to provide and obtain permission for using copyrighted works.[175] One example of how this might be helpful in the AI context is a company engaged in heavy research and development activities that may want to make additional internal uses of a large number of textual works that they've acquired lawfully.[176] The company may not have the bandwidth to engage in additional negotiations, while the publishers of the various scholarly journals would similarly be interested in licensing but would prefer to rely on a more streamlined approach. Importantly, voluntary collective licenses complement direct licenses, providing a framework where copyright owners and users can rely on collective licenses for many typical use cases and direct licenses for unique or individualized situations.

In the case of AI, we believe that both direct and collective licenses can be valuable to reduce uncertainty and establish a viable ecosystem going forward[177]. Some uses, such as certain training activities or general categories of outputs that need access to diffuse copyrighted materials, may be good candidates for

---

[172] *See* DANIEL GERVAISE, COLLECTIVE MANAGEMENT OF COPYRIGHT AND RELATED RIGHTS Ch. 1 (3rd ed. 2015).

[173] This is precisely what collective and centralized licensing does, namely allow users to use large repertoires of protected works. *See id.*

[174] *See* DANIEL GERVAIS, *The Economics of Copyright Collectives*, *in* 1 RESEARCH HANDBOOK ON THE ECONOMICS OF INTELLECTUAL PROPERTY LAW 489-507 (P. Menell & B. Depoorter eds, 2019).

[175] *See* Gervais, *supra* note 172.

[176] This is basically the fact pattern in Am. Geophysical Union v. Texaco, Inc., 60 F.3d 913, 926 (2d Cir. 1994). The court found that this activity was not a fair use.

[177] One example may be *Getty Images Launches Industry-First Model Release Supporting Data Privacy In Artificial Intelligence And Machine Learning*, GETTY IMAGES (Mar. 21, 2022), https://newsroom.gettyimages.com/en/getty-images/getty-images-launches-industry-first-model-release-supporting-data-privacy-in-artificial-intelligence-and-machine-learning.

collective licensing.[178] Conversely, specific high-value or individual uses based on more defined sets of copyrighted materials could be better suited for direct licensing.[179] Regardless of the approach, licensing provides both parties with compliant access to high-quality works, leading to innovative uses.

This is not unlike how licensing has worked during prior times of technological advancement.[180] In the 1970s, when photocopying was the disruptive technology, both direct and collective licensing helped make the market for using copyrighted materials work.[181] Similar stories exist about the early days of the Internet and the early days of TDM. Each technological advancement raised new questions, but licensing has always provided a long-standing answer to enable responsible and beneficial use of copyrighted works.

CONCLUSION

The history of copyright is one of constant adaptation to technological change.[182] When author's rights were first established, the underlying premise was to provide authors and their industry partners (publishers) with the means to live from the fruit of their labor by creating a viable marketplace for copies and later, for live public performances of music and theater. Since then, there have been a litany of developments from player pianos to the internet.

Placing all these adaptations on a timeline and then comparing them, it is clear that a key principle to ensure authors who have a say, or at least a right to be compensated, for new commercially significant uses of their works. Indeed, most commercial uses of copyright protected materials are subject to authors' rights, except in cases where a license is unlikely to be granted but there is a societal interest in allowing the use, such as parody.[183]

---

[178] *See e.g.*, Dave Shumaker, *A Work in Progress: CCC and Artificial Intelligence*, INFORMATION TODAY (Apr. 16, 2024), https://newsbreaks.infotoday.com/NewsBreaks/A-Work-in-Progress-CCC-and-Artificial-Intelligence-163574.asp.

[179] *See supra* note 177.

[180] *See infra* note 182.

[181] For example, *see Statement of Maria A. Pallante Register for Copy. of U.S. Copy. Off. Before the Subcomm. On Courts., Intellectual Prop. And Internet Comm. On the Judiciary,* U.S. Copy. Off. (Mar. 20, 2013) https://www.copyright.gov/regstat/2013/regstat03202013.html; *The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market-Legal Aspects* (Feb 2018), https://www.europarl.europa.eu/RegData/etudes/IDAN/2018/604941/IPOL_IDA(2018)604941_EN.pdf.

[182] For a short history, *see* Gervais, *supra* note 172. For a much more complete history of early copyright law in common law jurisdictions, *see* MARK ROSE, AUTHORS AND OWNERS: THE INVENTION OF COPYRIGHT (1995); and L. RAY PATTERSON, COPYRIGHT IN HISTORICAL PERSPECTIVE (1968).

[183] *Id.*

Now the most profound technological change in history is upon us.[184] A technology that can produce commercially competitive content that is likely to displace some human-created works. It can do this because it has absorbed the works of human authors.[185] The stakes could not be higher.

In this article, we reviewed the copyright aspects of LLM training and fine-tuning and concluded that copyright-relevant copying occurs during these processes. We have summarized the law of several jurisdictions and painted a varied picture of the scope of exceptions and limitations that may apply to LLM training, fine-tuning, and use (outputs), including several blurry areas of law that are likely to take years to be clarified by the courts. To reduce uncertainty for the benefit of all stakeholders, we examine the role that licensing solutions can play in this regard.

[184] Oddly, those who oppose any adaptation of the current framework sometimes say in the same breath that the current law is fine and that AI is too big a change for copyright to adapt to.

[185] *See supra* Part I(B).