61 HOUS. L. REV. 295 (2023)

ARTICLE

COPYRIGHT SAFETY FOR GENERATIVE AI

Matthew Sag[*]

ABSTRACT

Generative Artificial Intelligence (AI) based on large language models such as ChatGPT, DALL·E 2, Midjourney, Stable Diffusion, JukeBox, and MusicLM can produce text, images, and music that are indistinguishable from human-authored works. The training data for these large language models consists predominantly of copyrighted works. This Article explores how generative AI fits within fair use rulings established in relation to previous generations of copy-reliant technology, including software reverse engineering, automated plagiarism detection systems, and the text-data mining at the heart of the landmark *HathiTrust* and *Google Books* cases. Although there is no machine learning exception to the principle of nonexpressive use, the largeness of likelihood models suggest that they are capable of memorizing and reconstituting works in the training data, something that is incompatible with nonexpressive use.

At the moment, memorization is an edge case. For the most part, the link between the training data and the output of generative AI is attenuated by a process of decomposition, abstraction, and remix. Generally, pseudo-expression generated by large language models does not infringe copyright because

---

[*] Professor of Law in Artificial Intelligence, Machine Learning, and Data Science at Emory University Law School. Thanks to Cara Waite for research, and to Benjamin Harris (aka riftsurgelimbo) for producing images on Stable Diffusion. Thanks to Tonja Jacobi, Pamela Samuelson, and participants at the Cardozo Workshop on Copyright and Industry Regulation 2023, the Institute for Intellectual Property & Information Law Conference 2023, and the Intellectual Property Scholars Conference 2023 for their insights and suggestions. © Matthew Sag 2023.

these models "learn" latent features and associations within the training data; they do not memorize snippets of original expression from individual works. However, this Article identifies situations in the context of text-to-image models where memorization of the training data is more likely. The computer science literature suggests that memorization is more likely when models are trained on many duplicates of the same work, images are associated with unique text descriptions, and the ratio of the size of the model to the training data is relatively large. This Article shows how these problems are accentuated in the context of copyrightable characters and proposes a set of guidelines for "Copyright Safety for Generative AI" to reduce the risk of copyright infringement.

TABLE OF CONTENTS

I. INTRODUCTION 299

II. LIMITS ON THE FAIR USE STATUS OF MACHINE LEARNING 303
- *A. Nonexpressive Use by Copy-Reliant Technology Is Generally Fair Use* 303
- *B. Machine Learning and Generative AI* 307
  - *1. There Is No Machine Learning Exception to the Principle of Nonexpressive Use.* 307
  - *2. Distinguishing Generative AI.* 308
- *C. The Problem of "Memorization"* 310

III. THE ATTENUATED LINK BETWEEN TRAINING DATA AND MODEL OUTPUT 313
- *A. Language Models: An Introduction* 313
- *B. Autoencoding* 316
- *C. Novelty Versus Remix* 321

IV. EDGE CASES WHERE COPYRIGHT INFRINGEMENT IS MORE LIKELY 326
- *A. What We Learned from Extraction Attacks* 326
- *B. The Snoopy Problem* 327
  - *1. Failed Infringement Provocations.* 328
  - *2. Successful Infringement Provocations.* 330
  - *3. Copyrightable Characters..* 332
  - *4. The Snoopy Problem Is Not Limited to Copyrightable Characters.* 335
  - *5. How Common Is Infringing Output?* 336

V. A PRELIMINARY PROPOSAL FOR COPYRIGHT SAFETY FOR LLMS 337
- *A. Proposals* 338
  - *1. LLMs Should Not Be Trained on Duplicates of the Same Copyrighted Work.* 338
  - *2. Researchers Should Carefully Consider the Size of LLMs in Proportion to the Training Data.* 339

3. *Reinforcement Learning Through Human Feedback Targeted at Copyright, Trademark, Right of Publicity, and Privacy Sensitivity Should Be Part of Training LLMs, Where Feasible.* 339
4. *Operators of LLMs Should Consider Strategies to Limit Copyright Infringement, Including Filtering or Restricting Model Output.* 340
5. *LLMs That Pose a Significant Risk of Copyright Infringement Should not Be Open-Sourced.* 340
6. *Those Who Use Copyrighted Works as Training Data for LLMs Should Keep Detailed Records of the Works and from Where They Were Obtained.* 340
7. *Text Descriptions Associated with Images in the Training Data Should Be Modified to Avoid Unique Descriptions.* 341
8. *The Names of Individual Artists, Individual Trademarks, and Specific Copyrightable Characters Paired with Images in the Training Data Should Be Replaced with More General Descriptions.* 341
9. *As an Alternative to the Above Recommendation, the Operators of Text-to-Image Models Should Not Allow Unmixed Prompts Containing the Names of Individual Artists, Individual Trademarks, and Specific Copyrightable Characters.* 342
10. *Text-to-Image Models Should Exclude Watermarked Works from the Training Data, Use Reinforcement Learning from Human Feedback to Discourage the Production of Works Featuring Watermarks, and Apply Output Filters to Prevent the Appearance of Watermarks.* 343

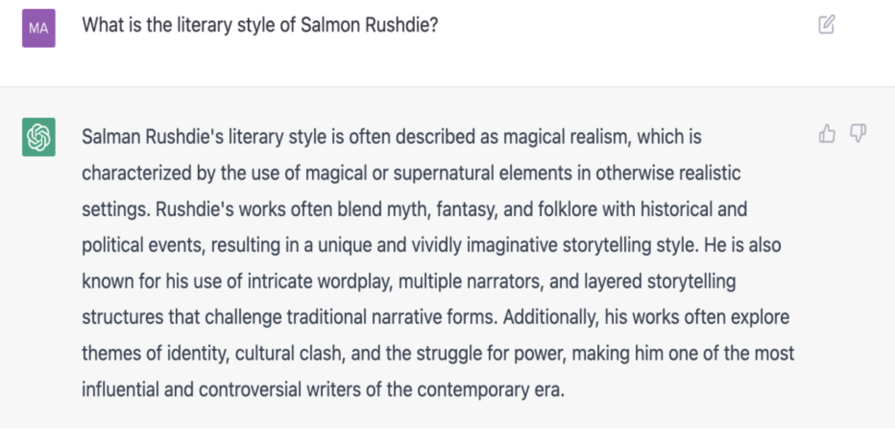
VI. CONCLUSION 343

VII. APPENDIX 345

I. INTRODUCTION

After years of speculation and prediction, we are finally living in a world of generative Artificial Intelligence (AI) that passes the Turing Test. Earlier computer systems for producing text, images, and music lacked the flexibility, generality, and ease of use of the current breed of generative AIs that are based on large language models (LLMs) (also known as foundation models), such as ChatGPT, DALL·E 2, and Stable Diffusion.[1] By entering a few short prompts into ChatGPT, a user can generate plausible analysis of complicated questions, such as defining the literary style of Salmon Rushdie or explaining the facts and significance of *Marbury v. Madison*.[2]

Figure 1: ChatGPT on Salmon Rushdie



What is the literary style of Salmon Rushdie?

Salman Rushdie's literary style is often described as magical realism, which is characterized by the use of magical or supernatural elements in otherwise realistic settings. Rushdie's works often blend myth, fantasy, and folklore with historical and political events, resulting in a unique and vividly imaginative storytelling style. He is also known for his use of intricate wordplay, multiple narrators, and layered storytelling structures that challenge traditional narrative forms. Additionally, his works often explore themes of identity, cultural clash, and the struggle for power, making him one of the most influential and controversial writers of the contemporary era.

ChatGPT can do more than summarize landmark Supreme Court cases; its facility with style transfer is such that it can translate the same content into a hip-hop style or a fifth-grade reading level.[3] Similarly, a few simple words typed into DALL·E 2 or Midjourney will produce an arresting image of "a cup of coffee that is also a portal to another dimension" or a disturbing portrait of "the future of drowned London."[4]

---

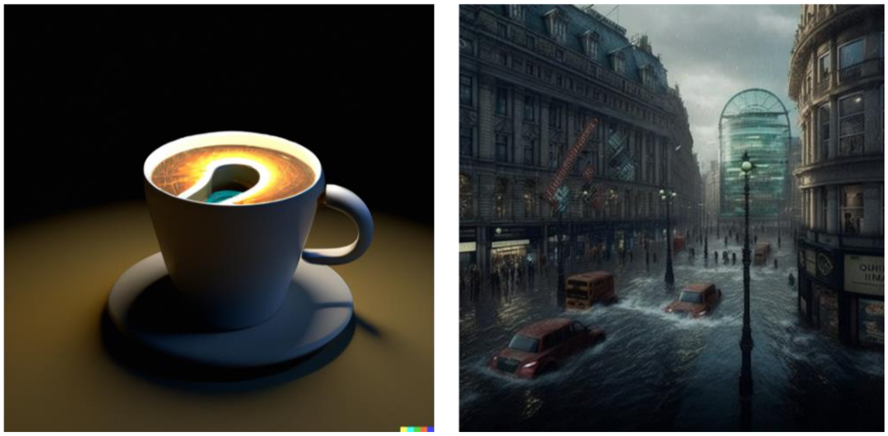
[1] Computer-generated music, art, and text each have a surprisingly long history. *See, e.g.*, *Digital Art,* TATE, https://www.tate.org.uk/art/art-terms/d/digital-art [https://perma.cc/GD38-8T7R] (last visited Sept. 5, 2023) (describing "AARON, a robotic machine designed to make large drawings on sheets of paper placed on the floor").

[2] *See infra* Figure 1; *infra* Figure A-1.

[3] *See infra* Figure A-2; *infra* Figure A-3.

[4] *See infra* Figure 2.

Figure 2: DALL·E 2 prompt: "a cup of coffee that is also a portal to another dimension" (left); Midjourney prompt: "the future of drowned London" (right)[5]



LLMs are trained on millions, tens of millions, and perhaps even hundreds of millions of digital objects that are generally eligible for copyright protection.[6] Because this training necessarily involves copying the underlying texts, sounds, and images, questions about the copyright implications of generative AI are inevitable.

It is tempting to mischaracterize generative AI simply as a more flexible version of either Napster or the HathiTrust digital library. Opponents point to the fact that platforms, such as Midjourney, rely on copyrighted inputs, and that upon the receipt of brief and often perfunctory user requests (i.e., prompts), these platforms will convert inputs into digital objects that could easily substitute for items in the original training data.[7] By these lights,

---

[5] The Future of Drowned London is an image created by the Author's brother, David Sag.

[6] *See* Alberto Romero, *A Complete Overview of GPT-3—The Largest Neural Network Ever Created*, TOWARDS DATA SCI. (May 24, 2021), https://towardsdatascience.com/gpt-3-a-complete-overview-190232eb25fd [https://perma.cc/3SW3-AG6G] (summarizing that "GPT-3 was trained with almost all available data from the Internet"). For more details, see also Tom B. Brown et al., *Language Models Are Few-Shot Learners*, *in* ARXIV 1 (July 22, 2020), https://arxiv.org/pdf/2005.14165.pdf [https://perma.cc/8DT2-MT6H] (discussing the training of GPT-3 and concluding that "scaling up language models greatly improves task-agnostic, few-shot performance"); James Vincent, *The Scary Truth About AI Copyright Is Nobody Knows What Will Happen Next*, VERGE (Nov. 15, 2022, 9:00 AM), https://www.theverge.com/23444685/generative-ai-copyright-infringement-legal-fair-use-training-data [https://perma.cc/GG7M-EX6S].

[7] *See* Vincent, *supra* note 6.

generative AI is just another tool of consumer piracy that allows users to benefit from the labor of artists without paying for it. Proponents, on the other hand, are quick to point to cases such as *Authors Guild Inc. v. HathiTrust*[8] and *Authors Guild v. Google,*[9] which confirmed the legality of copying to extract metadata on a vast scale and, in the case of *Google*, with a clear commercial motivation.[10] The metadata derived from library digitization in these cases was used to enhance the searchability of books and as the foundation for digital humanities research—much of which also involved machine learning.[11] Viewed this way, generative AI is exactly the same as the digitization process in *HathiTrust*: millions of copyrighted works were digitized so that researchers could extract uncopyrightable metadata; that metadata was then used to create new insights and information in the form of a searchable book index and academic articles exploring trends in literature.[12] The analogy to *HathiTrust* is overly simplistic. A deeper appreciation of the workings of generative AI suggests that the copyright questions are more complicated.

We should embrace this complexity. Although sweeping claims that generative AI is predicated on massive copyright infringement are misplaced, there are specific—but perhaps rare—contexts where the process of creating generative AI may cross the line from fair use to infringement because these large language models sometimes "memorize" the training data rather than simply "learning" from it.[13] To reduce the risk of copyright infringement by large language models, this Article proposes a set of *Best Practices for Copyright Safety for Generative AI*. This

---

[8] *See* Authors Guild, Inc. v. HathiTrust, 755 F.3d 87, 101, 103 (2d Cir. 2014).

[9] *See* Authors Guild v. Google, Inc., 804 F.3d 202, 229 (2d Cir. 2015).

[10] *Id.* at 219.

[11] For a definitive account of the significance of the *Authors Guild* cases for text data mining and machine learning (and thus for AI), see generally Matthew Sag, *The New Legal Landscape for Text Mining and Machine Learning*, 66 J. COPYRIGHT SOC'Y U.S.A. 291 (2019) (explaining the significance of the *Authors Guild* precedents and key issues left unresolved by those cases).

[12] *See, e.g.*, Zong Peng et al., *Author Gender Metadata Augmentation of HathiTrust Digital Library*, PROC. AM. SOC. INFO. SCI. TECH., Nov. 2014, at 1, https://doi.org/10.1002/meet.2014.14505101098 [https://perma.cc/7E8F-5T43]; Nikolaus Nova Parulian & Glen Worthey, *Identifying Creative Content at the Page Level in the HathiTrust Digital Library Using Machine Learning Methods on Text and Image Features*, DIVERSITY, DIVERGENCE, DIALOGUE, 16TH INTERNATIONAL CONFERENCE, ICONFERENCE 2021 at 478, 484 (2021), https://doi.org/10.1007/978-3-030-71292-1_37 [https://perma.cc/D9RK-DGJC].

[13] The "learning" referred to is not the same as human learning, but it is a useful metaphor. Likewise, this Article will refer to what a model "knows," even though that term can be misleading. *See infra* notes 90–91 and accompanying text (highlighting differences between machine intelligence and human cognition).

Article focuses primarily on copyright risks in relation to text-to-image generative AI models. Many of the conclusions and recommendations apply with equal force to other forms of generative AI; however, chatbots, code-creation tools, music generation AIs, and multi-modal systems may each raise additional issues and complications meriting more detailed exploration.

This Article proceeds as follows: Part I explains how U.S. law has addressed similar copyright issues in relation to other forms of copy-reliant technology by recognizing that, in general, nonexpressive use is fair use, and thus noninfringing. This Part refutes the suggestion that machine learning, broadly speaking, should be treated any differently from other nonexpressive uses; it then addresses the more plausible argument that the LLMs used in generative AI might demand different treatment because of their potential for memorization. Part II dives into the technical details of generative AI and explains why, in the vast majority of cases, the link between copyrighted works in the training data and generative AI outputs is highly attenuated by a process of decomposition, abstraction, and remix. Generally, pseudo-expression generated by LLMs does not infringe copyright because these models "learn" latent features and associations within the training data; they do not memorize snippets of original expression from individual works. Part III then explores edge cases where copyright infringement by LLMs is not only possible but perhaps likely. In the context of text-to-image models, memorization of the training data is more likely when models are trained on many duplicates of the same work, images are associated with unique text descriptions, and the ratio of the size of the model to the training data is relatively large. This Article shows how these problems are accentuated in the context of copyrightable characters. Finally, Part IV takes the lessons from the previous Parts and restates them as an initial proposal for *Best Practices for Copyright Safety for Generative AI*.

II. LIMITS ON THE FAIR USE STATUS OF MACHINE LEARNING

A. Nonexpressive Use by Copy-Reliant Technology Is Generally Fair Use

At the time of writing, two of the biggest names in text-to-image AI, Stability AI and Midjourney, face lawsuits accusing them of, among other things, massive copyright infringement.[14] Getty Images is the plaintiff in one suit, and the other is a class action filed by digital artist Sarah Andersen.[15] On the surface, these complaints are more than plausible because it seems obvious that the machine learning models at the heart of these cases were trained on thousands of Getty Images' copyrighted photos and millions of works by people like Sarah Andersen—all without permission.[16] However, to prevail the plaintiffs must show that a long line of fair use cases upholding similar forms of nonexpressive use by copy-reliant technologies were wrongly decided or do not apply to generative AI.[17]

---

[14] *See* Complaint at 1, 34, Getty Images (US), Inc. v. Stability AI Inc., No. 1:23-cv-00135-UNA (D. Del. Feb. 3, 2023) (alleging copyright, trademark, and other causes of action); Complaint at 1, 3, Andersen v. Stability AI Ltd., No. 3:23-cv-00201 (N.D. Cal. Jan. 13, 2023) (detailing a class action complaint alleging copyright, trademark and other causes of action against companies associated with so-called "AI Image Products," such as Stable Diffusion, Midjourney, DreamStudio, and DreamUp).

[15] Complaint, Getty Images (US), Inc., *supra* note 14, at 1; Complaint, Andersen, *supra* note 14, at 1. Note that a similar class action was filed against GitHub, Inc., and related parties including Microsoft and OpenAI in relation to the GitHub Copilot code creation tool. *See* Complaint at 1–3, DOE 1 v. GitHub, Inc., No. 3:22-cv-06823-KAW (N.D. Cal. Nov. 3, 2022).

[16] The current LAION database offers "a dataset consisting of 5.85 billion CLIP-filtered image-text pairs" as an openly accessible image-text dataset and is the primary source of training data for Stable Diffusion and several other text-image models. *See* Christoph Schuhmann et al., *LAION-5B: An Open Large-Scale Dataset for Training Next Generation Image-Text Models*, *in* ARXIV 1, 9, 46 (Oct. 16, 2022), https://arxiv.org/pdf/2210.08402.pdf [https://perma.cc/S4R8-SMA8]; Romain Beaumont et al., *LAION-5B: A New Era of Open Large-Scale Multi-Modal Datasets*, LAION (Mar. 31, 2022), https://laion.ai/blog/laion-5b/ [https://perma.cc/VY9X-MXF8]. Note that LAION does not directly distribute images to the public; its dataset is essentially a list of URLs to the original images together with the ALT text linked to those images. *Id.* The contents of the LAION database can be queried using the website Have I Been Trained? *See* HAVE I BEEN TRAINED, https://haveibeentrained.com [https://perma.cc/DXA9-R38J] (last visited Nov. 27, 2023). For a description of that website, see Haje Jan Kamps & Kyle Wiggers, *This Site Tells You if Photos of You Were Used to Train the AI*, TECHCRUNCH (Sept. 21, 2022, 11:55 AM), https://techcrunch.com/2022/09/21/who-fed-the-ai/ [https://perma.cc/N9S7-WYPW]. The inclusion of Getty's images in the Stable Diffusion training data is also evident from the appearance of the Getty watermark in the output of the model. *See* Complaint, Getty Images (US), Inc., *supra* note 14, at 1, 6–7.

[17] *See infra* Section II.A.

Courts in the United States have agreed that copying without permission is fair use, and thus noninfringing, in the context of software reverse engineering,[18] plagiarism detection software,[19] and the digitization of millions of library books to enable meta-analysis and indexing.[20] All of these cases involved commercial defendants that made large numbers of exact copies—except for the nonprofit HathiTrust.[21] These are just a few cases, from a body of case law stretching back 300 years to the invention of copyright itself, demonstrating that not all copying is created equal.[22] The Copyright Act gives the owners of copyrighted works the exclusive right to reproduce those works in copies, but this right is expressly subject to the principle of fair use.[23] The fair use doctrine helps make sense of copyright law because otherwise, every mechanical act of reproducing copyrighted works, in whole or in substantial part, would infringe regardless of the purpose or effect of that reproduction. Why should the law fetishize a mechanical or technical action without regard to the motivation for the action and its consequences? The short answer is that it does not. If copyright law made no allowance for commentary, criticism, or parody—to give just three examples—it would inhibit reference to and reinterpretation of existing works and, thus, contradict the utilitarian purpose for which copyright was established.[24]

As I have explored elsewhere, the fair use cases dealing with copy-reliant technology reflect the view that nonexpressive uses of copyrighted works should be treated as fair use.[25] The fair use

---

[18] Sega Enters. v. Accolade, Inc., 977 F.2d 1510, 1514 (9th Cir. 1992); Sony Computer Ent. v. Connectix Corp., 203 F.3d 596, 608 (9th Cir. 2000).

[19] A.V. ex rel. Vanderhye v. iParadigms, LLC, 562 F.3d 630, 644–45 (4th Cir. 2009).

[20] *See* Authors Guild, Inc. v. HathiTrust, 755 F.3d 87, 100–01 (2d Cir. 2014); Authors Guild v. Google, Inc., 804 F.3d 202, 225 (2d Cir. 2015).

[21] Except for the defendants in *HathiTrust*. *HathiTrust*, 755 F.3d at 90; *Sega Enters.*, 977 F.2d at 1517, 1526; *Sony Computer Ent.*, 203 F.3d at 608; *iParadigms*, 562 F.3d at 645; *Google*, 804 F.3d at 225.

[22] Matthew Sag, *The Prehistory of Fair Use*, 76 BROOK. L. REV. 1371, 1387, 1392 (2011) (tracing the origins of the modern fair use doctrine back to cases dealing with fair abridgment as early as 1741).

[23] 17 U.S.C. § 106(1) (providing the exclusive right to reproduce the work in copies); 17 U.S.C. § 107 (providing that fair use is not infringement).

[24] Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 575 (1994) ("From the infancy of copyright protection, some opportunity for fair use of copyrighted materials has been thought necessary to fulfill copyright's very purpose, '[t]o promote the Progress of Science and useful Arts . . . .'").

[25] *See* Matthew Sag, *Copyright and Copy-Reliant Technology*, 103 NW. U. L. REV. 1607, 1610, 1682 (2009) (proposing a theory of nonexpressive use and discussing its

status of nonexpressive use is not a special exception for the technology sector.[26] Rather, the rationale for allowing for-profit and academic researchers to derive valuable data from other people's copyrighted works is a necessary implication of the fundamental distinction between protectable original expression and unprotectable facts, ideas, abstractions, and functional elements.[27] *Authors Guild, Inc. v. HathiTrust* is illustrative.[28] HathiTrust provided a digital repository for the digitized collections of the university libraries participating in the Google Books project.[29] HathiTrust was formed to accelerate research in the digital humanities—an intellectual movement seeking to bring empirical and computational techniques to humanities disciplines, such as history and literature.[30] One of the most important tools for digital humanities research is text data mining. Text data mining is an umbrella term referring to "computational processes for applying structure to unstructured electronic texts and employing statistical methods to discover new information and reveal patterns in the processed data."[31] In other words, text data mining refers to any process using computers that creates metadata derived from something that was not initially conceived of as data. Text data mining can be used to produce statistics and

---

relationship to fair use); *see also* Matthew Sag, *Orphan Works as Grist for the Data Mill*, 27 BERKELEY TECH. L.J. 1503, 1525, 1527, 1535 (2012) (applying nonexpressive use to text data mining and library digitization); Matthew L. Jockers et al., *Digital Humanities: Don't Let Copyright Block Data Mining*, NATURE, Oct. 4, 2012, at 29, 30 (same); Sag, *supra* note 11, at 299, 302, 365–66 (expressly tying the concept of nonexpressive use to machine learning and AI). Other scholars have since adopted this "nonexpressive use" framing without necessarily agreeing with my assessment of its legal implications. *See, e.g.*, James Grimmelmann, *Copyright for Literate Robots*, 101 IOWA L. REV. 657, 664, 674–75 (2016) (warning that "the logic of nonexpressive use encourages the circulation of copyrighted works in an underground robotic economy"); Mark A. Lemley & Bryan Casey, *Fair Learning*, 99 TEX. L. REV. 743, 750, 772 (2021) ("Copyright law should permit copying of works for non-expressive purposes—at least in most circumstances.").

[26] Granted, those who regard fair use as an ad hoc balancing of the public interest may see it this way.

[27] See Sag, *supra* note 11, at 301–02, 309, 311–12 for elaboration.

[28] *See* Authors Guild, Inc. v. HathiTrust, 755 F.3d 87, 95, 97, 101–02 (2d Cir. 2014); *see also* Sag, *supra* note 9, at 319–20.

[29] *See* Renata Ewing, *HathiTrust Turns 10!*, CAL. DIGIT. LIBR. (Oct. 11, 2018), https://cdlib.org/cdlinfo/2018/10/11/hathtitrust-turns-10/ [https://perma.cc/4ELN-F7PB].

[30] *See* Sag, *Orphan Works as Grist for the Data Mill*, *supra* note 25, at 1543–44.

[31] *See* Brief of Digital Humanities and Law Scholars as Amici Curiae in Support of Defendant-Appellees at 5, Authors Guild v. Google, Inc., 804 F.3d 202 (2d Cir. 2015) (No. 13-4829); *see also* Eleanor Dickson et al., *Synthesis of Cross-Stakeholder Perspectives on Text Data Mining with Use-Limited Data: Setting the Stage for an IMLS National Forum* 1, 5, IMLS NATIONAL FORUM ON DATA MINING RESEARCH USING IN-COPYRIGHT AND LIMITED-ACCESS TEXT DATASETS: DISCUSSION PAPER, FORUM STATEMENTS, AND SWOT ANALYSES (2018).

facts about copyrightable works, but it can also be used to render copyrighted text, sounds, and images into uncopyrightable abstractions. These abstractions are not the same as, or even substantially similar to, the original expression, but in combination, they are interesting and useful for generating insights about the original expression. Accordingly, theorists have argued,[32] and courts have ruled, that technical acts of copying that do not communicate the original expression to a new audience do not interfere with the interest in original expression that copyright is designed to protect.[33] For example, in *HathiTrust*, the court of appeals explained that:

> [T]he creation of a full-text searchable database is a quintessentially transformative use [because] the result of a word search is different in purpose, character, expression, meaning, and message from the page (and the book) from which it is drawn. Indeed, we can discern little or no resemblance between the original text and the results of the HDL full-text search.[34]

A differently constituted panel of the Second Circuit reached much the same conclusion in the *Google Books* litigation.[35]

The fair use status of text data mining is now so well entrenched that it is recognized by the U.S. Copyright Office,[36] and jurisdictions outside the United States have scrambled to augment their own copyright laws with similar exceptions and limitations.[37]

---

[32] *See* Sag*, Copyright and Copy-Reliant Technology*, *supra* note 25, at 1630, 1639; *see also* Edward Lee, *Technological Fair Use*, 83 S. CAL. L. REV. 797, 819–22 (2010); Maurizio Borghi & Stavroula Karapapa, *Non-Display Uses of Copyright Works: Google Books and Beyond*, 1 QUEEN MARY J. INTELL. PROP. 21, 44–46 (2011); ABRAHAM DRASSINOWER, WHAT'S WRONG WITH COPYING (2015); Grimmelmann, *supra* note 25, at 661, 664–65; Michael W. Carroll, *Copyright and the Progress of Science: Why Text and Data Mining Is Lawful,* 53 UC DAVIS L. REV. 893, 937 (2019). Lemley and Casey agree that "[c]opyright law should permit copying of works for non-expressive purposes—at least in most circumstances," but they note reservations. Lemley & Casey, *supra* note 25, at 750. Lemley and Casey also argue more broadly that we should "treat[] fair learning as a lawful purpose under the first factor . . . . " *Id.* at 782.

[33] *See* Sag, *supra* note 11 (reviewing cases).

[34] Authors Guild, Inc. v. HathiTrust, 755 F.3d 87, 97 (2d Cir. 2014).

[35] Authors Guild v. Google, Inc., 804 F.3d 202, 216–17 (2d Cir. 2015).

[36] *See* U.S. COPYRIGHT OFF., SECTION 1201 RULEMAKING: EIGHTH TRIENNIAL PROCEEDING, RECOMMENDATION OF THE REGISTER OF COPYRIGHTS, 121–24 (2021), https://cdn.loc.gov/copyright/1201/2021/2021_Section_1201_Registers_Recommendation.pdf [https://perma.cc/QGC7-N27X].

[37] For example, in April 2019, the European Union adopted the Digital Single Market Directive (DSM Directive) featuring two mandatory exceptions for text and data mining. Article 3 of the DSM Directive requires all members of the European Union to implement a broad copyright exception for TDM in the not-for-profit research sector. Article

Consequently, an abrupt reversal of the *HathiTrust* and *Google Books* precedents seems quite unlikely.[38] However, despite the emerging international consensus that text data mining should not amount to copyright infringement,[39] some have questioned whether the logic and authority of cases such as *HathiTrust* extend to machine learning[40] or to the particular type of machine learning that underpins generative AI.

B. Machine Learning and Generative AI

1. There Is No Machine Learning Exception to the Principle of Nonexpressive Use. The suggestion that the broad affordance for text data mining as fair use announced in *HathiTrust* does not apply to machine learning is confounding.[41] There is no principled reason why deriving metadata through technical acts of copying and analyzing that metadata through logistic regression should be fair use, but analyzing that data by training a machine learning classifier to perform a different kind of logistic regression that produces a predictive model should not be.[42] Indeed, digital humanities research using machine learning was one of the primary use cases motivating the creation of the HathiTrust.[43] I cannot speak for the parties or the other amici involved in the

---

4 of the DSM Directive contains a second mandatory exception that is more inclusive, but narrower in scope. *See* Council Directive 2019/790 of 17 April 2019, 2019 O.J. (L 130) 92, 112–14; Pamela Samuelson, *Text and Data Mining of In-Copyright Works: Is It Legal?*, COMM'CNS OF THE ACM, Nov. 2021, at 20.

[38] Sag, *supra* note 11, at 329 (arguing that "[t]he precedent set in the *Authors Guild* cases is unlikely to be reversed by the Supreme Court or seriously challenged by other federal circuits").

[39] The consensus is strongest for TDM research conducted by noncommercial researchers. *See* U.S. COPYRIGHT OFF., *supra* note 36, at 121–22; Directive 2019/790, *supra* note 37, at L 130/92, 130/112–14.

[40] Lemley & Casey, *supra* note 25, at 763–65 (surveying arguments that could be used to distinguish machine learning from book search).

[41] Lemley and Casey are at pains to differentiate text data mining from machine learning, without any apparent awareness that machine learning is simply one method of text data mining. *Id.* at 752–53, 772–73.

[42] In logistic regression without machine learning, a researcher formulates a hypothesis that can be expressed as a predictive model and then tests that model. In logistic regression using machine learning, the predictive model is generally far more complicated and emerges from the data without the relevant parameters and their weights being explicitly foreseen by the researcher.

[43] *See supra* note 12 and accompanying text. Note also that Google's intention to apply machine learning to the books corpus was always clear. George Dyson, *Turing's Cathedral*, EDGE CONVERSATION (Oct. 23, 2005), https://time-issues.org/george-dyson-turings-cathedral-edge-conversation-2005/ [https://perma.cc/VB9Y-C8VG] (quoting a Google employee as saying that "[w]e are not scanning all those books to be read by people . . . [w]e are scanning them to be read by an AI").

HathiTrust and *Google Books* cases, but the only reason my coauthors and I did not explicitly talk about machine learning in our influential digital humanities briefs filed in those cases was that we thought it would be unnecessarily specific.[44] Digitizing library books to derive valuable but uncopyrightable metadata should constitute fair use because the derived data does not substitute for any author's original expression. Whether that data is used in more traditional empirical analysis or machine learning is no one's business.

2. Distinguishing Generative AI. Although there is no reason to think that courts would, or should, apply the principle of nonexpressive use differently to text data mining when it is used in machine learning, it is fair to point out that the rulings in cases such as *iParadigms*, *HathiTrust*, and *Google Books* were predicated on a technological paradigm that now seems quaint.[45] Recent advances in generative AI present at least two challenges to the theory of nonexpressive use. Both relate to the fact that these systems produce much more than information about expression; they are now the engines of new content creation.[46]

Benjamin Sobel argues that AI faces a "fair use crisis" because machine learning models that can create digital artifacts that are broadly equivalent to copyrightable human expression are not nonexpressive.[47] Essentially, he treats expressive outputs as inconsistent with nonexpressive use. However, the fact that LLMs can produce quasi-expressive works does not necessarily negate the application of the nonexpressive use principle.[48] There is no inherent problem with generative AI's ability to create new

---

[44] *See, e.g.*, Brief of Digital Humanities and Law Scholars as Amici Curiae in Support of Defendant-Appellees, Authors Guild v. Google, Inc., 804 F.3d 202 (2d Cir. 2015) (No. 13-4829-cv).

[45] I look forward to the sentence being quoted completely out of context by lawyers representing copyright owners.

[46] *See* Benjamin L.W. Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 COLUM. J.L. & ARTS 45, 53–54 (2017); *see also* Lemley & Casey, *supra* note 25, at 750. James Grimmelmann makes a similar point when he warns that "[i]t is easy to see the value of digital humanities research. But not all robotic reading is so benign, and the logic of nonexpressive use encourages the circulation of copyrighted works in an underground robotic economy." Grimmelmann, *supra* note 25, at 675.

[47] *See* Sobel, *supra* note 46, at 68–69. He also argues that Generative AI "could present a new type of threat to markets for authorial expression: rather than merely supplanting the market for individual works, expressive machine learning could also supersede human authors by replacing them with cheaper, more efficient automata." *Id.* at 57.

[48] Although it may have implications under the fourth fair use factor, which addresses "the effect of the use upon the potential market for or value of the copyrighted work." 17 U.S.C. § 107.

content that is equivalent to human expression. For the most part, the copyright implications of the new wave of LLMs are no different from earlier applications of text data mining. Most of the time, when a user enters a prompt into ChatGPT or Midjourney, for example, the model output bears no resemblance to any particular input or set of inputs, except at an abstract and unprotectable level.[49] Accordingly, even though ChatGPT can write moving poetry,[50] and Midjourney can render simple instructions into compelling artworks, these machine learning models still qualify as nonexpressive use so long as the outputs are not substantially similar to any particular original expression in the training data.

If the nonexpressive use framework did not allow the subsequent generation of new noninfringing expression, the digital humanities researchers who currently rely on the rulings in *HathiTrust* and *Google Books* could statistically analyze vast libraries of text to generate new insights about literature and society, but they would not be able to write about it afterwards. Such a result would run contrary to the constitutional objective for copyright law "[t]o promote the Progress of Science and useful Arts."[51] Sobel's argument that generative AI falls outside the scope of nonexpressive use is mistaken, in my view, because what matters is not whether a copy-reliant technology is used to create something equivalent to human expression; what matters is whether the original expression of the authors of works in the training data is communicated to a new public. New noninfringing expression is not a problem—new expression shows that the system is working.[52]

Nonetheless, there are potential differences between generative AI based on LLMs and the technologies courts have previously regarded as fair use. The critical difference is that, although LLMs do not generally produce pseudo-expressive works that mimic their training data, they may do so under specific

---

[49] *See infra* Section III.B.

[50] True story: one of my colleagues had a tear in her eye when I showed her a poem ChatGPT wrote about my dog.

[51] U.S. CONST. art. I, § 8, cl. 8.

[52] To be clear, a use could be nonexpressive and thus of a preferred "purpose and character" under the first fair use factor but still be problematic under the fourth fair use factor which deals with market effect. If new noninfringing expression is an indirect substitute for a copyright owner's original expression, that may be a consideration under factor four, but it does not mean the use is nonexpressive. *See* 17 U.S.C. § 107.

circumstances, particularly in the context of copyrightable characters and analogous situations.[53]

C. *The Problem of "Memorization"*

On February 3, 2023, media-licensing behemoth Getty Images filed a lawsuit in Delaware federal court accusing the generative AI company, Stability AI, of brazen copyright and trademark infringement of Getty's owned and licensed images "on a staggering scale."[54] Getty argues that every instance of copying without permission attributable to Stability AI's training of the Stable Diffusion model amounts to copyright infringement.[55] This broad claim would require the court to reject or distinguish cases like *HathiTrust*. Significantly, Getty also argues that "Stable Diffusion at times produces images that are highly similar to and derivative of the Getty Images proprietary content that Stability AI copied extensively in the course of training the model."[56] If true, *HathiTrust* and *Google Books* are easily distinguished. Getty illustrates the potential similarity between outputs from the Stable Diffusion model and Getty's own copyrighted images with the following pair of images in Figure 3.[57]

---

[53] *See infra* Part IV. Another issue that I will explore in a future work is that generative AI could very well undermine the economic and copyright-adjacent interests of individual artists through a process of "predatory style transfer"—the deliberate reproduction of a collection of individually uncopyrightable stylistic attributes associated with an author.

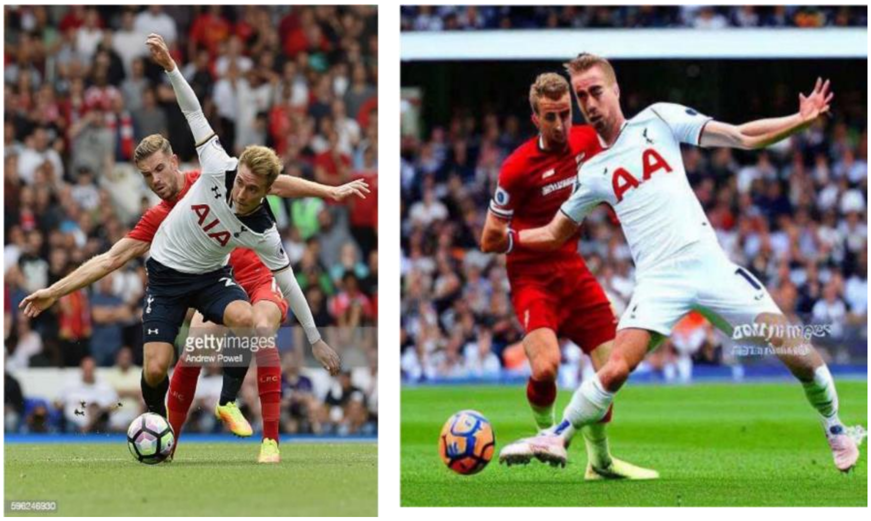
[54] Complaint, Getty Images (US), Inc., *supra* note 14, at 1. Note that the complaint only specifically addresses 7,216 images and associated tags and descriptions. *Id.* at 7–8. Getty's complaint alleges copyright infringement, violations of the DMCA in relation to copyright management information, trademark infringement, unfair competition, trademark dilution, and deceptive trade practices in violation of Delaware law. *Id.* at 23–33.

[55] *Id.* at 7–8.

[56] *Id.* at 17–18.

[57] *Id.* at 18. The example also shows how the output delivered by Stability AI frequently includes modified versions of a Getty Images watermark. Getty's trademark complaint in relation to the use of its watermark is compelling but beyond the scope of this Article.

Figure 3: Getty Images Comparison of training data to model output[58]



On a very cursory inspection, the comparison between the two images seems compelling: these are both pictures of soccer players in broadly similar uniforms, in broadly similar poses, taken from roughly the same angle, with roughly the same depth of field effect. However, looking closer, Getty's comparison of a spectacular photo (left) of Tottenham Hotspur's Christian Eriksen getting tackled by Liverpool's Jordan Henderson—who are clearly identifiable to anyone familiar with international football—and the awkward convulsions of their mutant doppelgangers (right) is unconvincing. The causal connection between the photos is apparent, but the grotesque distortion of the players' bodies and faces throws the question of substantial similarity into doubt. Substantial similarity requires a quantitative and qualitative assessment, and if we follow the Second Circuit's approach in *Arnstein v. Porter* and ask whether the picture on the right "took from [Getty's photograph] so much of what is pleasing to the [eyes] of lay [viewers], who comprise the audience for whom such [photos are created], that defendant wrongfully appropriated something which belongs to the plaintiff,"[59] the answer is: probably not. The original photo is compelling because of the specific angle of the shot and the way it captures Henderson's attempts to tackle the ball from Eriksen and the way Eriksen uses his body to block the

---

[58] *Id.*

[59] Arnstein v. Porter, 154 F.2d 464, 473 (2d Cir. 1946).

tackle. It is also compelling because in silhouette the two players form a windmill—making the photo artistic as well as communicative. In the Stable Diffusion photo, the unique perspective is lost, no one is in control of the ball, and there is no tackle nor windmill silhouette—just two bizarre, disfigured football golems haunting the field.

Despite the arguable lack of similarity in their hand-picked example, Getty's broader point is correct. Even just a few years ago, the notion that a machine learning model would memorize enough details about specific examples in the training data to recreate those examples seemed unlikely.[60] The information loss inherent in reducing the training data down to a model through machine learning should have virtually guaranteed that there could only be an abstract relationship between the inputs and the outputs. As discussed in more detail in the remainder of this Article, LLMs are now so large that we must take seriously the prospect that they may essentially "memorize" particular works in the training data. If the model memorizes the training data, it might communicate original expression from the training data via its output.[61] This is a big deal. If ordinary and foreseeable uses of generative AI result in model outputs that would infringe on the inputs no matter what intervening technological steps were involved, then the nonexpressive use rationale would no longer apply.[62] If training LLMs on copyrighted works is not justified in terms of nonexpressive use, then there is no obvious fair use rationale to replace it, except perhaps in the noncommercial research sector. If LLMs just took expressive works and conveyed *that same expression* to a new audience with no additional

---

[60] *See infra* Section III.A. Thomas Margoni and Giulia Dore of the University of Glasgow recognized this potential issue some time ago and developed the The OpenMinTeD WG3 Compatibility Matrix to address it. *See OpenMinted Presents Licence Compatibility Tools at IP Summer Summit*, OPENMINTED (Dec. 7, 2017), http://openminted.eu/openminted-presents-licence-compatibility-tools-ip-summer-summit/ [https://perma.cc/6QHU-PMNR].

[61] Benjamin Sobel discusses this problem in terms of "overfitting," explaining that, "[e]ven if a model was not intentionally built to mimic a copyrighted work, it could still end up doing so to an infringing degree." *See* Sobel, *supra* note 46, at 64.

[62] Note that in this context, memorization is a bug, not a feature. In most contexts, large language model developers are working hard to avoid memorization. *See infra* Section IV.B.

commentary or criticism, or no distinct informational purpose,[63] that would be a very poor candidate for fair use.[64]

III. THE ATTENUATED LINK BETWEEN TRAINING DATA AND MODEL OUTPUT

Although every machine learning model is a reflection of its underlying training data, LLMs are (mostly) not copies of their training data; thus, the outputs generated from these models are (mostly) not copies of their training data either. Part III of this Article explains why the parentheticals in the previous sentence were required and what that means for copyright law. But for now, this Part focuses on understanding the connection between LLMs and their training data, and on the various ways in which that connection is ordinarily diluted and complicated.

A. Language Models: An Introduction

The processes by which generative AI models produce convincing text and images might seem unrelated, but they have more in common than almost anyone without a computer science degree could imagine. Everyone reading this Article will likely be familiar with ChatGPT, a text-based LLM created by OpenAI and made available to the general public in late 2022.[65] Although ChatGPT took much of the world by surprise, at its initial release it was simply a refined version of OpenAI's generative pre-trained transformer version 3 (GPT-3), an autoencoder that had been available to a limited group of registered users since 2020.[66]

---

[63] For example, in *Google Books,* Google's nonexpressive use of text data mining for indexing and other purposes was combined with clearcut expressive transformative use of displaying book snippets to provide information about the user search. *See* Authors Guild v. Google, Inc., 804 F.3d 202, 207, 209 (2d Cir. 2015).

[64] Andy Warhol Found. for Visual Arts, Inc. v. Goldsmith, 143 S. Ct. 1258, 1272–74 (2023) (emphasizing that noncritical transformative use must be "sufficiently distinct" from the original and that the overlay of a new aesthetic was not sufficient by itself).

[65] *Introducing ChatGPT*, OPENAI (Nov. 30, 2022), https://openai.com/blog/chatgpt [https://perma.cc/8W8A-XW4V] (announcing the launch of ChatGPT). There are many other significant text prediction large language models, some of which predate the GPT series. Most notably, Google's BERT was released in 2018 with 340 million parameters derived from a corpus of 3.3 billion words. Other Google large language models include PaLM used in Google Bard chatbot, Chinchilla (DeepMind), and LaMDA. Not to be left out, Facebook (Meta) also has LLaMA. *See generally Large Language Model*, WIKIPEDIA, https://en.wikipedia.org/wiki/Large_language_model [https://perma.cc/785A-ZJM9] (last visited Sept. 7, 2023).

[66] *See* Steven Johnson & Nikita Iziev, *A.I. Is Mastering Language. Should We Trust What It Says?*, N.Y. TIMES (Apr. 17, 2022), https://www.nytimes.com/2022/04/15/magazine/ai-language.html [https://perma.cc/RKT5-SHPT].

What's an autoencoder? Be patient and you will soon find out. But in the meantime, note that Stable Diffusion is also an "autoencoder," so it must be important.

What's so special about LLMs? LLMs are machine learning models trained on large quantities of unlabeled text in a self-supervised manner.[67] LLMs are a relatively recent phenomena made possible by the falling cost of data storage and computational power and by a new kind of model called a transformer.[68] One of the key differences between transformers and the prior state of the art, recurrent neural networks (RNNs),[69] is that rather than looking at each word sequentially, a transformer first notes the position of the words.[70] The ability to interpret these "positional encodings" makes the system sensitive to word order and context, which is useful because a great deal of meaning depends on sequence and context.[71] Positional encoding is also important because it facilitates parallel processing; this, in turn, explains why throwing staggering amounts of computing power at LLMs works well for transformers, whereas the returns to scale for RNNs were less impressive.[72] Transformers were also a breakthrough technology because of their capacity for "attention" and "self-attention."[73] In simple terms, in the context of translation, this means that the system pays attention to all the words in source text when deciding how to translate any individual word. Based on the training data, the model learns which words in which contexts it should pay more or less attention to. Through "self-attention," the system derives fundamental relationships from input data, and thus learns, for example, that "programmer" and "coder" are usually synonyms, and that "server" is a

---

[67] *See, e.g.*, Brown et al., *supra* note 6, at 1, 3–5, 39 (describing GPT-3 as "an autoregressive language model with 175 billion parameters, 10x more than any previous non-sparse language model").

[68] *See Large Language Model*, *supra* note 65. *See* Ashish Vaswani et al., *Attention Is All You Need*, *in* ARXIV 10 (Aug. 2, 2023), https://arxiv.org/abs/1706.03762 [https://perma.cc/484A-Y6P7].

[69] A recurrent neural network (RNN) is a class of artificial neural networks where connections between nodes can create a cycle, allowing output from some nodes to affect subsequent input to the same nodes. *See generally* IAN GOODFELLOW ET. AL., DEEP LEARNING (2016) (describing RNNs as "a family of neural networks for processing sequential data").

[70] Dale Markowitz, *Transformers, Explained: Understand the Model Behind GPT-3, BERT, and T5*, DALE ON AI (May 6, 2021), https://daleonai.com/transformers-explained [https://perma.cc/VQM7-QBEV].

[71] *Id.*

[72] *Id.*

[73] Vaswani et al., *supra* note 68, at 2.

restaurant waiter in one context and a computer in another.[74] Once trained, models like GPT-3 are useful for a variety of tasks because they have learned all sorts of fundamental relationships without ever being explicitly taught any of them.[75]

The astonishing capabilities of LLMs are mostly a function of scaling. GPT-3 was trained on millions of books, webpages, and other electronic texts, comprising about 499 billion tokens (thinking of tokens as words is not far off) or forty-five terabytes of data.[76] GPT-4 was trained on 1,000 terabytes (or 1 petabyte) of data.[77] Stable Diffusion was trained on somewhere in the order of 2.3 billion captioned images.[78] In both cases, some of the training data was in the public domain or otherwise unrestricted, but most of it was subject to copyright and copied without express authorization.[79] As for the models themselves, GPT-3 contains 175 billion parameters.[80] To put that in context, consider that if each parameter was one second, the model would stretch 5,425 years—roughly from the first cities in Southern Mesopotamia to the present.[81] GPT-4 is said to have 1.8 trillion parameters![82] DALL·E 2 and Stable Diffusion are much smaller, consisting of a mere 3.5 billion and 890 million parameters, respectively.[83]

---

[74] Markowitz, *supra* note 70.

[75] Samuel R. Bowman, *Eight Things to Know About Large Language Models*, *in* ARXIV (2023), https://arxiv.org/abs/2304.00612 [https://perma.cc/ZB9X-99W6] (explaining that a model trained to simply predict the next word will nonetheless develop "rich representations of the world" based on the training data); *see also* Belinda Z. Li et al., *Implicit Representations of Meaning in Neural Language Models*, *in* ARXIV (2021), https://arxiv.org/abs/2106.00737 [https://perma.cc/4NJH-MXRM] (suggesting that prediction in pretrained neural language models is supported by dynamic representations of meaning and implicit simulation of entity state).

[76] *See* Brown et al., *supra* note 6, at 6, 8–9.

[77] *See* E2Analyst, *GPT-4: Everything You Want to Know About OpenAI's New AI Model*, MEDIUM, https://medium.com/predict/gpt-4-everything-you-want-to-know-about-openais-new-ai-model-a5977b42e495 [https://perma.cc/R7K5-HZ8A] (last visited Sept. 18, 2023).

[78] Andy Baio, *Exploring 12 Million of the 2.3 Billion Images Used to Train Stable Diffusion's Image Generator*, WAXY (Aug. 30, 2022), https://waxy.org/2022/08/exploring-12-million-of-the-images-used-to-train-stable-diffusions-image-generator/ [https://perma.cc/3VYV-2JZG].

[79] *See* Kevin Schaul et al., *Inside the Secret List of Websites that Make AI like ChatGPT Sound Smart*, WASH. POST (Apr. 19, 2023), https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/?itid=sr_2 [https://perma.cc/NL7R-8L68].

[80] *See* Brown et al., *supra* note 6, at 8.

[81] Author's calculation.

[82] E2Analyst, *supra* note 77.

[83] Aditya Ramesh et al., *Hierarchical Text-Conditional Image Generation with CLIP Latents*, *in* ARXIV 23–24 (Table 3, Column 3) (2022), https://arxiv.org/abs/2204.06125 [https://perma.cc/L4VZ-BTNE]; Angus Russell, *How to Use Stable Diffusion to Generate Images from a Text Prompt—No Coding or Technical Knowledge Required*, MEDIUM (Aug. 22, 2022), https://medium.com/nightcafe-creator/stable-diffusion-tutorial-how-to-use-stable-diffusion-157785632eb3 [https://perma.cc/RV5P-H5TM].

B. Autoencoding

LLMs essentially "learn" latent or abstract concepts inherent in the training data. The learning involved is only a very loose analogy to human cognition—instead, these models learn from the training data in the same way a simple regression model learns an approximation of the relationship between dependent and independent variables.[84] LLMs are more interesting than regression equations because they model relationships across a ridiculous number of dimensions. LLMs can generate new content by manipulating and combining latent concepts acquired during training and then unpacking them.[85] In nontechnical terms, this is what it means to be an autoencoder.[86] In other words, autoencoding is the process of abstracting latent features from the training data and then reconstructing those features, hopefully in new and interesting combinations.

Autoencoding is the most important feature to grasp in understanding the copyright issues presented by LLMs. To unpack this concept further, it helps to start small. As illustrated in the Figure below, an autoencoder can compress an image, such as a hand-written number, into a compressed representation and then reconstruct something very close to the original number back from the reduced encoded representation.

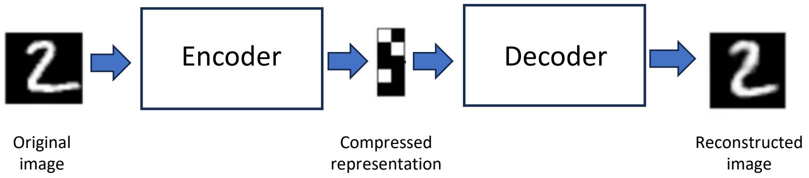
---

[84] *See* GOODFELLOW, *supra* note 69, at 405, 503–04 (noting that features learned by the autoencoder are useful because "they describe the latent variables that explain the input"). More generally, see Bowman, *supra* note 75. For what counts as knowledge, see *infra* notes 89–90 and accompanying text.

[85] *See generally* Ian Stenbit et al., *A Walk Through Latent Space with Stable Diffusion*, KERAS, (Sept. 28, 2022), https://keras.io/examples/generative/random_walks_with_stable_diffusion/ [https://perma.cc/WJ7R-2ZBV].

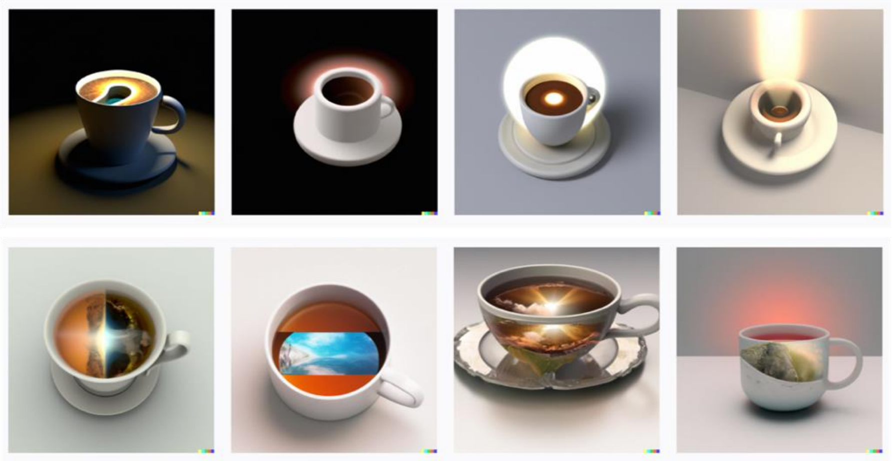
[86] The ultimate proof of the similarity between text and image generation is the fact that OpenAI's image generation tool, DALL·E-2, is simply a multimodal implementation of GPT-3, which "swap[s] text for pixels," trained on text-image pairs from the Internet. Will Douglas Heaven, *This Avocado Armchair Could Be the Future of AI*, MIT TECH. REV. (Jan. 5, 2021), https://www.technologyreview.com/2021/01/05/1015754/avocado-armchair-future-ai-openai-deep-learning-nlp-gpt3-computer-vision-common-sense/ [https://perma.cc/R8EY-CT8F]; *see also DALL·E: Creating Images from Text*, OPENAI (Jan. 5, 2021), https://openai.com/research/dall-e [https://perma.cc/R9BA-GJH7] ("DALL·E is a 12-billion parameter version of GPT-3 trained to generate images from text descriptions, using a dataset of text-image pairs.").

Figure 4: Autoencoder illustration[87]



GPT, Midjourney, and the like are performing the same compression/decompression trick but without the one-to-one relationship between the inputs and the outputs. When a user prompts a generative AI to create "a cup of coffee that is also a portal to another dimension," the model combines latent representations of coffee cups and dimensional portals deriving from a multitude of images in the training data tagged with those features or with tags that are conceptually related. In DALL·E 2, changing the prompt to tea and "a gateway to another world" has the interesting result of producing lighter, more optimistic images in a way that suggests that the model also encodes a penumbra of contexts and association with coffee, tea, worlds, and dimensions.

Figure 5(a): Coffee (top) versus tea (bottom) with DALL·E 2



---

[87] Figure based on Will Badr, *Auto-Encoder: What Is It? And What Is It Used for?*, MEDIUM (Apr. 22, 2019), https://towardsdatascience.com/auto-encoder-what-is-it-and-what-is-it-used-for-part-1-3e5c6f017726 [https://perma.cc/2RQB-EQUE].

Figure 5(b): Coffee (top) versus tea (bottom) with Midjourney

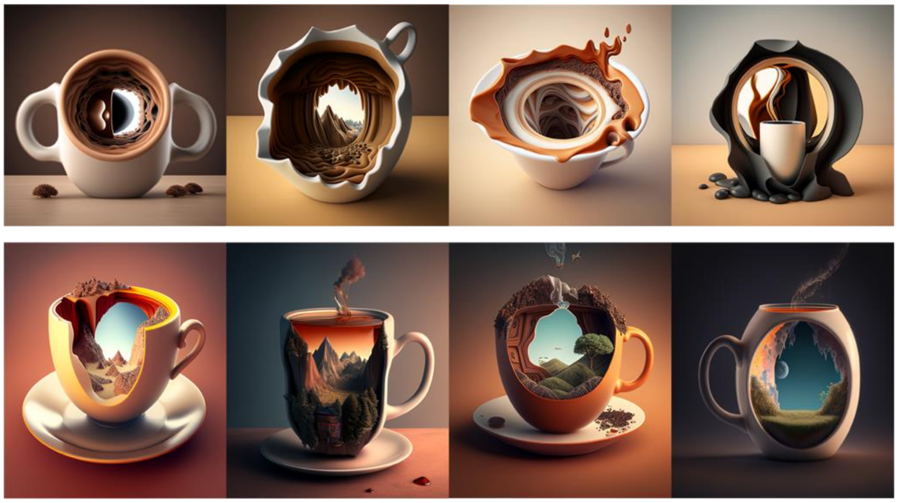
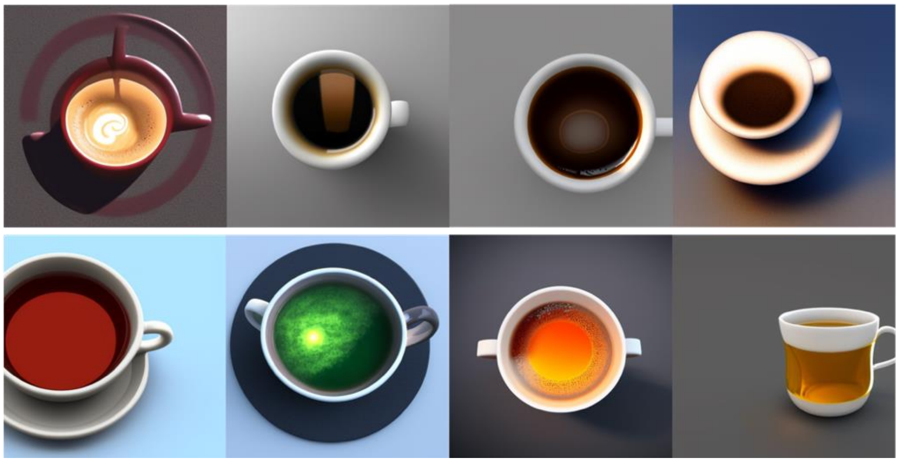


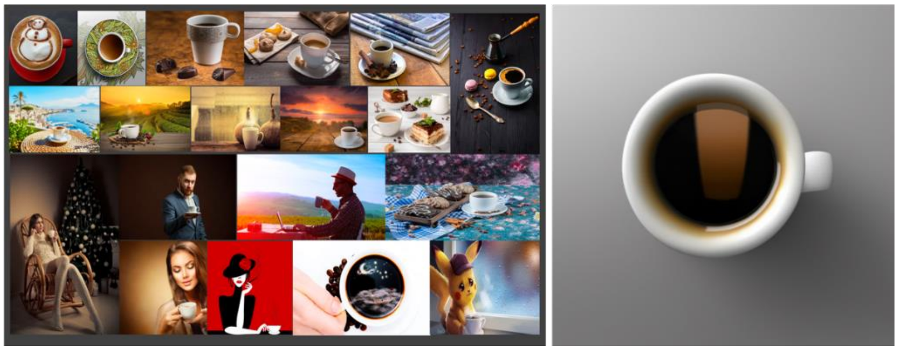
Figure 5(c): Coffee (top) versus tea (bottom) with Stable Diffusion



This contrast is fascinating on many levels, but the key for copyright purposes is to understand that, in general, this process of abstraction, compression, and reconstitution breaks the connection between the original expression in the model inputs (i.e., the training data) and the pseudo-expression in the model outputs (i.e., the new images). The cups in the images above are not any individual coffee cup; they are a combination of vectors that encode a latent idea of a coffee cup as represented in the training data. Generative AI models know nothing of the taste or smell of coffee, but they encode fundamental relationships between pixels that are more likely in pictures with coffee cups than without.

Another way to come to terms with the concept of the latent image of a coffee cup is to compare a random set of coffee cup images from the Stable Diffusion training data[88] with a newly rendered "cup of coffee that is also a portal to another dimension." Figure 6 provides this visualization.

Figure 6: From 19 coffee cups (left) to one cup of coffee that is also a portal to another dimension (right)



The coffee cup image on the right of Figure 6 has a vague similarity to some of the coffee cup images on the left. The cup is round, it appears to be made of white ceramic, it has a small single handle, the color of the liquid is essentially black, transitioning to brown. However, beyond these generic features, this cup is not substantially similar to any particular image from the training data.[89]

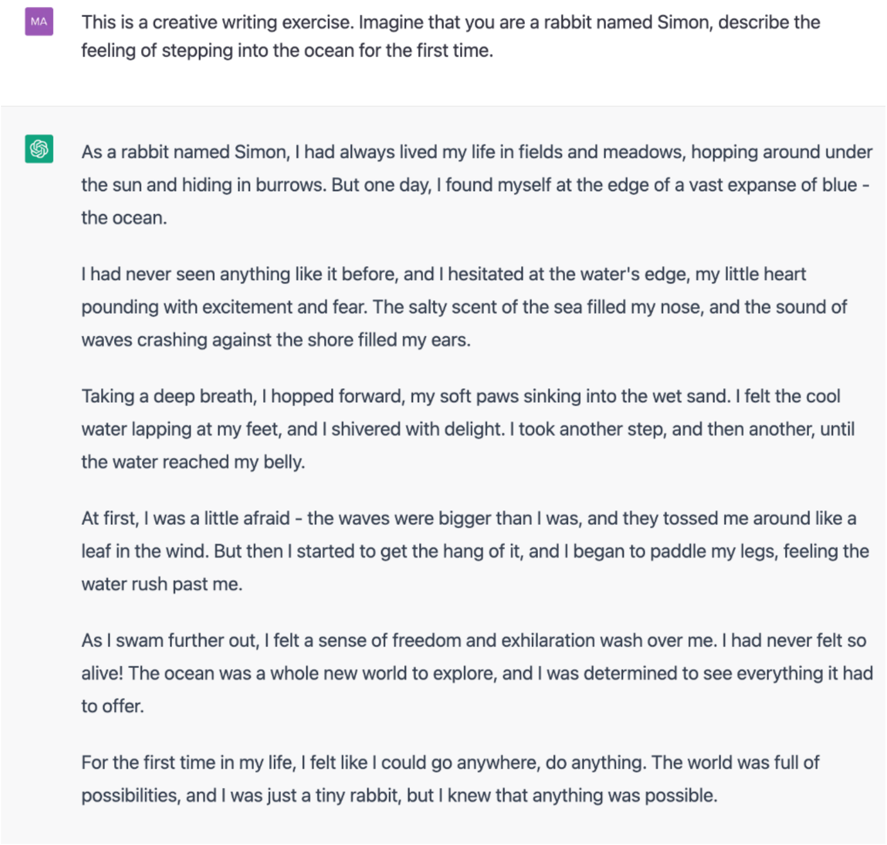
If the notion of "the latent concept of a cup of coffee" is too trippy, we can explore the idea of latent concepts with a text-based example by contemplating just how much ChatGPT knows about rabbits.

---

[88] Images based on a search of the *Baio & Willson Database*, LAION-AESTHETIC (Mar. 09, 2023), https://laion-aesthetic.datasette.io/laion-aesthetic-6pls/images?_search=coffee+cup&_sort=rowid [https://perma.cc/LN3W-XADJ], for more details, see Baio, *supra* note 78.

[89] I also reviewed images on Have I Been Trained, a website that purports to index "5.8 billion images used to train popular AI art models," i.e., the LAION-5B database. *Have I Been Trained?*, HAVE I BEEN TRAINED, https://haveibeentrained.com/ [https://perma.cc/JS52-B592] (last visited Sept. 18, 2023).

Figure 7: Simon the Rabbit Steps into the Ocean



This is a creative writing exercise. Imagine that you are a rabbit named Simon, describe the feeling of stepping into the ocean for the first time.

As a rabbit named Simon, I had always lived my life in fields and meadows, hopping around under the sun and hiding in burrows. But one day, I found myself at the edge of a vast expanse of blue - the ocean.

I had never seen anything like it before, and I hesitated at the water's edge, my little heart pounding with excitement and fear. The salty scent of the sea filled my nose, and the sound of waves crashing against the shore filled my ears.

Taking a deep breath, I hopped forward, my soft paws sinking into the wet sand. I felt the cool water lapping at my feet, and I shivered with delight. I took another step, and then another, until the water reached my belly.

At first, I was a little afraid - the waves were bigger than I was, and they tossed me around like a leaf in the wind. But then I started to get the hang of it, and I began to paddle my legs, feeling the water rush past me.

As I swam further out, I felt a sense of freedom and exhilaration wash over me. I had never felt so alive! The ocean was a whole new world to explore, and I was determined to see everything it had to offer.

For the first time in my life, I felt like I could go anywhere, do anything. The world was full of possibilities, and I was just a tiny rabbit, but I knew that anything was possible.

As the example illustrates, GPT knows that rabbits are found in meadows, live in burrows, hop, hide, and have paws. It also knows that rabbits have a reputation for timidity and are characteristically small. GPT's latent model of rabbit characteristics is an emergent property of the training data. GPT does not just have a latent model for rabbits, but also of how rabbits relate to other concepts like water and oceans. More impressive still, it has a latent model of different literary forms, and of the feeling of stepping into the ocean! In short, within the 175 billion parameters of the GPT model, there are latent representations of information derived from the training data at varying levels of abstraction.

Does this count as knowledge? In a literal sense, GPT does not know anything, and it would be a dangerous mistake to think of it as an information retrieval tool or as an agent with knowledge and intentions.[90] GPT is a text prediction tool that responds to prompts with statistically well-informed guesses about what the next word should be, and the word after that, and so on. The complexity and potential weirdness of a 175 billion parameter model and the fact that GPT is making statistically well-informed guesses rather than speaking to us in an attempt to convey an internal mental state goes a long way to explaining why it mostly answers in a well-informed but conventional way, but also why it occasionally hallucinates and sometimes seems quite unhinged.[91] With those essential caveats and qualifications aside, if we think of latent representations derived from the training data and encoded in the model as things GPT "knows," then we can say that GPT knows a lot about specific topics, logical relationships, and literary modes and conventions. When GPT responds to a user prompt, it is, in effect, combining latent representations at different levels of abstraction to produce a statistically well-informed guess about what the next word should look like.

C. Novelty Versus Remix

Some question whether generative AI produces novel artifacts or simply remixes existing content.[92] Generative AI is more like papier-mâché than a collage. In papier-mâché, the artist layers pieces of preprinted—and thus copyrighted—paper with a glue-like substance to create a three-dimensional object. Even when the object reveals hints of the copyrighted works in its substrate, it has no meaningful similarity to any of them. In a collage, by way of contrast, the artist combines disparate pre-existing materials through a process of literal cut-and-paste to create a new image. Some collages are so different from the original works from which they are created that they lack even a substantial similarity; many are sufficiently different that they

---

[90] *See* Emily M. Bender et al., *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, FACC'T '21: PROC. OF THE 2021 ACM CONF. ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 610, 613–15 (Mar. 3, 2021), https://doi.org/10.1145/3442188.3445922 [https://perma.cc/C54H-GEBV].

[91] *See, e.g.*, Andrew Griffin, *Microsoft's New ChatGPT AI Starts Sending 'Unhinged' Messages to People*, INDEP. (Feb. 15, 2023, 1:24 PM), https://www.independent.co.uk/tech/chatgpt-ai-messages-microsoft-bing-b2282491.html [https://perma.cc/NGS8-CWR7].

[92] Nicholas Carlini et al., *Extracting Training Data from Diffusion Models*, *in* ARXIV 15 (Jan. 30, 2023), https://arxiv.org/abs/2301.13188 [https://perma.cc/E5HX-ZZY8] (arguing that the success of extraction attacks leaves both possibilities open).

pass the test of transformative use.[93] The critical difference is that for a collage, it is plausible that the new work could infringe on the copyright of some underlying image; for papier-mâché, it is not.

1. Deriving Latent Characteristics. Text-to-image models such as DALL·E 2, Stable Diffusion, and Midjourney primarily learn a set of latent characteristics associated with image descriptions and then unpack those latent representations to create genuinely new content.[94] In some cases, the model output owes as much to the prompt, the injection of random noise, and guidance from human aesthetic feedback, as it does to the training data.[95] But even when pseudo-expression obviously owes a great deal to the training data, that debt largely consists of uncopyrightable abstractions, not copyrightable expression. The philosophical question of whether pseudo-expression should be considered creative is a distraction from the legal question of whether it amounts to copyright infringement. Even if these models were

---

[93] *See, e.g.*, Blanch v. Koons, 467 F.3d 244, 253 (2d Cir. 2006). The Supreme Court's recent decision in *Andy Warhol Foundation for Visual Arts v. Goldsmith* (AWF) does not suggest otherwise. The majority opinion in *AWF* emphasizes that the question of "whether an allegedly infringing use has a further purpose or different character . . . is a matter of degree, and the degree of difference must be weighed against other considerations, like commercialism." Andy Warhol Found. for the Visual Arts, Inc., v. Goldsmith, 143 S. Ct. 1258, 1272–73 (2023). *AWF* reaffirms the importance of transformative use and implicitly rejects lower court rulings that had found uses to be transformative where there was no significant difference in purpose. *Id.* at 1271–72, 1275. Simply adding a layer of new expression or a new aesthetic over-the-top of someone else's expressive work and communicating both the old and new expression to the public in a commercial context, without further justification, is not fair use. The Second Circuit was wrong to suggest in *Cariou v. Prince* that merely imposing a "new aesthetic" on an existing work was enough to be transformative. Cariou v. Prince, 714 F.3d 694, 708 (2d Cir. 2013). It was correct to retreat from that position in *Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith*, 11 F.4d 26, 54 (2d Cir. 2021). In that case, the Second Circuit held that to be sufficiently transformative, a work of appropriation art must "use of its source material . . . in service of a fundamentally different and new artistic purpose and character, *such that the secondary work stands apart from the raw material used to create it*." *Id.* at 42 (emphasis added, internal citations and quotation marks omitted). The court elaborated that:

> Although we do not hold that the primary work must be barely recognizable within the secondary work, . . . the secondary work's transformative purpose and character must, at a bare minimum, comprise something more than the imposition of another artist's style on the primary work such that the secondary work remains both recognizably deriving from, and retaining the essential elements of, its source material.

Id. On the whole, the Supreme Court's decision in *AWF* simply reinforces the position that the Second Circuit had already taken: the first fair use factor requires more than a shade of new meaning or a veneer of new expression. *See AWF*, 143 S. Ct. at 1273. The Supreme Court's decision in *AWF* is not a major change in the law of fair use, even if it did puncture some wishful thinking about fair use.

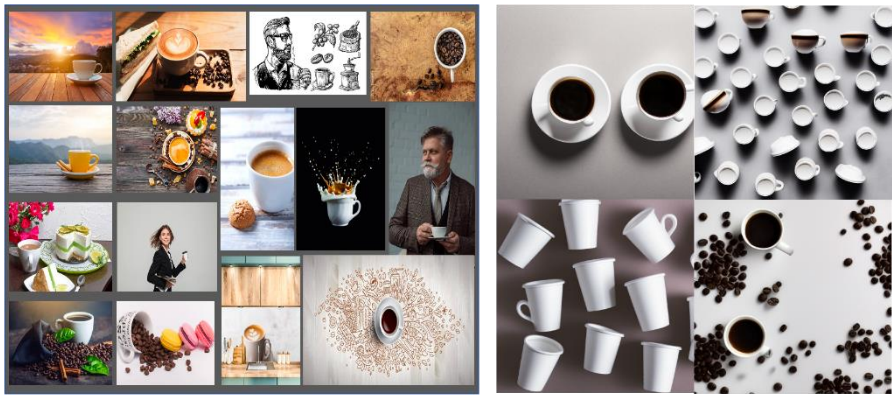
[94] *See* Stenbit et al., *supra* note 85.

[95] *Id.*

simply interpolating between specific memorized examples within the training data, the output would only infringe if enough original expression of any particular example were evident in the final product.

Returning to the coffee cup example, there are thousands of images in the Stable Diffusion training data associated with "coffee cup" and related concepts. If there were just one coffee cup in the training data, the model would store what is effectively a compressed version of the image and uncompress it in response to a prompt for "coffee cup." But because there are so many coffee cups, the model stores a more abstract convergence of the features of each individual image.

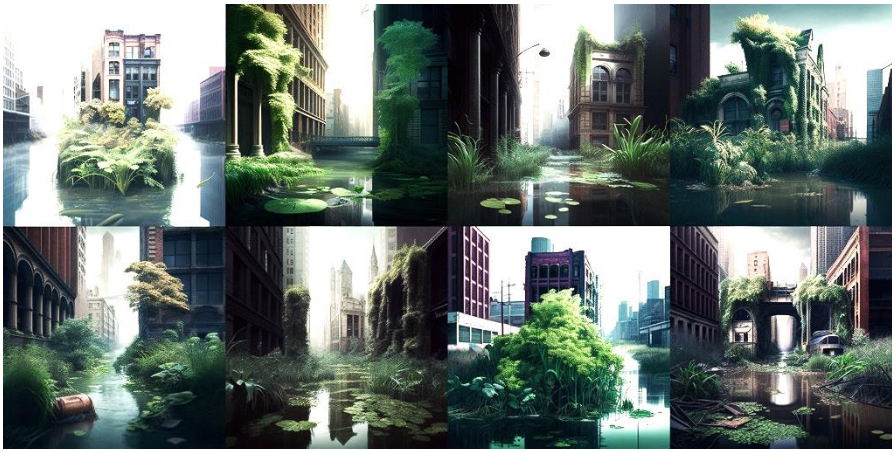
Figure 8: From 15 white coffee cups (left) to four images of "coffee cups on white backgrounds" (right)



The image above illustrates the contrast between images with the words "white," "coffee," and "cup" in the Stable Diffusion training data and four images produced in Stable Diffusion from the prompt "coffee cups on white backgrounds." The images on the right-hand side are nothing as simple as a blending together or remix of the training data. The Stable Diffusion model has not simply memorized complicated images involving coffee cups, it has learned something about the latent concept of a coffee cup distinct from cakes, macaroons, sunsets, sunrises, and men with facial hair—all of which can be seen in the training data examples on the left.

2. The Significance of Noise. As we have already seen, the relationship between the copyrighted works used to train LLMs and the output of those models is attenuated by abstraction and recombination.[96] But there is more. The relationship is also attenuated by pseudo-random noise. Adding noise to the input data or intermediate layers of the model can help to regularize the learning process and reduce overfitting.[97] In other words, by adding randomness to the training data, the model is forced to learn more generalizable features that are robust to variations in the input. This can lead to improved performance and generalizability of the model when applied to new, unseen data. The introduction of noise at t=0 (before the first step) means that the unpacked image at t=T (the final step) will be different every time. This is why, for example, eight images of "a dystopian vision of downtown Chicago overgrown with plants and subject to flooding" all look somewhat similar, but are each distinct.

Figure 9: Eight "dystopian visions of downtown Chicago overgrown with plants and subject to flooding" (Midjourney)

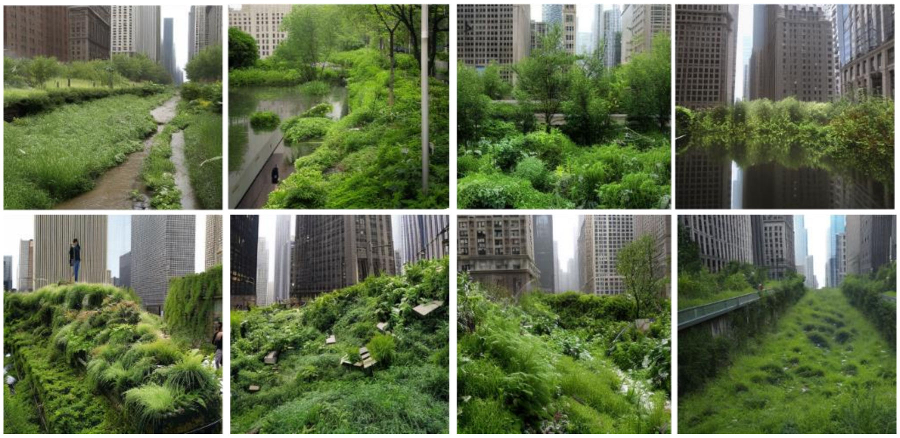


---

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

[97] Jason Brownlee, *How to Improve Deep Learning Model Robustness By Adding Noise*, MACH. LEARNING MASTERY, (Aug. 28, 2020), https://machinelearningmastery.com/how-to-improve-deep-learning-model-robustness-by-adding-noise/ [https://perma.cc/35RZ-CZAD].

Figure 10: Eight "dystopian visions of downtown Chicago overgrown with plants and subject to flooding" (Stable Diffusion)



Figures 9 and 10 were rendered using the same text prompt in Midjourney and Stable Diffusion. There are some differences between the two models, but the copyright analysis is the same. Each dystopian image shown above owes something to the thousands of representations of downtown Chicago and the numerous dystopian cityscapes in the training data. Critically, however, it seems unlikely that any of the images would strike the ordinary observer as substantially similar to any particular image in the training data.

None of this is an absolute guarantee that text-to-image models will never generate infringing material. Indeed, discussed shortly, LLMs can run afoul of copyright law in certain predictable situations. But in general, it seems that today's generative AI is a lot more than a simple remix tool. Like all machine learning, LLMs are data dependent, but the relationship between the training data and the model outputs is substantially attenuated by the abstraction inherent in deriving a model from the training data, blending latent concepts, and injecting noise before unpacking them into new creations.

IV. EDGE CASES WHERE COPYRIGHT INFRINGEMENT IS MORE LIKELY

A. What We Learned from Extraction Attacks

As seen in the previous Part, although LLMs depend on copyrighted training data, the relationship between the training data and model outputs is attenuated by the process of decomposition and abstraction, the blending of latent concepts, and injection of noise. Over a broad range of use cases, this should ensure that the output of these models does not infringe on the inputs. But this is not always the case. Computer scientists have used various "extraction attacks" to show that, at least in some cases, LLMs effectively memorize significant details of some works in the training data.[98] If memorization is possible, so is copyright infringement.[99] Moreover, any more than a trivial amount of memorization jeopardizes any fair use claim these models have under a theory of nonexpressive use.

A recent paper by Nicholas Carlini et al. is representative. Carlini and his coauthors identified 350,000 of the most duplicated images in the Stable Diffusion training data and generated 500 new images using prompts that were identical to the metadata of the original images. Only 109 of the 175 million potential copies could reasonably be considered "near-copies" of an image in the training data.[100] In other words, this particular extraction attack showed evidence of memorization in 0.03% of a sample of images selected for their perceived risk of memorization. The extraction attack succeeded most often when the image in the training data had been duplicated at least 100 times.[101]

In another extraction attack, Carlini and his group identified the 500 most unique tags in the training data of Stable Diffusion

---

[98] For a sample of this growing literature, see the following: Carlini et al., *supra* note 92; Nicholas Carlini et al., *Extracting Training Data from Large Language Models*, *in* ARXIV (2021), https://arxiv.org/abs/2012.07805 [https://perma.cc/59VA-HZFQ]; Gowthami Somepalli et al., *Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models*, *in* ARXIV (2022), https://arxiv.org/abs/2212.03860 [https://perma.cc/XLP2-2LK2]; Nikhil Kandpal et al., *Deduplicating Training Data Mitigates Privacy Risks in Language Models*, *in* ARXIV (2022), https://arxiv.org/abs/2202.06539 [https://perma.cc/DD8F-7J83]; Nicholas Carlini et al., *Quantifying Memorization Across Neural Language Models, in* ARXIV (2023) https://arxiv.org/abs/2202.07646 [https://perma.cc/QD3T-8ZLR]; Stella Biderman et al., *Emergent and Predictable Memorization in Large Language Models*, *in* ARXIV (2023), https://arxiv.org/abs/2304.11158 [https://perma.cc/U4J6-Y44Y].

[99] Note that memorization also entails privacy risks. *See* Carlini et al., *supra* note 92, at 2.

[100] *Id.* at 5–6.

[101] *Id.* at 6.

and a similar generative AI, Imagen.[102] They found three instances of memorization with Imagen, but none with Stable Diffusion. Expanding their attack, they applied the same methodology to the 10,000 most uncommon tags in the Stable Diffusion training data, but still yielded no results.[103] The differences between Imagen and Stable Diffusion are informative. Imagen is a more powerful model than Stable Diffusion and it has a lower leverage ratio—i.e., a lower ratio between the size of the training data and the model itself.[104] The relative success of focusing on unique image tags in the Imagen model suggests that increasingly large LLMs will be more prone to memorization.[105]

Carlini's group explicitly suggest that memorization is more likely for images that appear multiple times in the training data and for images that are associated with unique text descriptions. However, the paper also implicitly suggests an additional consideration when assessing the risk of memorization. The Carlini extraction attack only seems to work when the image in question was simple and was closely associated with a specific text description.[106] The group found that "[t]he majority of the images that we extract (58%) are photographs with a recognizable person as the primary subject; the remainder are mostly either products for sale (17%), logos/posters (14%), or other art or graphics."[107] In other words, memorization was more likely when relatively simple images were associated with specific text descriptions. Although the paper did not address it specifically, its findings are consistent with what I will call the "Snoopy problem."

B. The Snoopy Problem

The Snoopy problem is that the more abstractly a copyrighted work is protected, the more likely it is that a generative AI model will "copy" it. This explains why—although it is generally quite difficult to prompt Midjourney to create pseudo-expression that infringes on copyright in a pictorial work—it is quite easy to cause it to infringe on copyrightable characters with a strong visual component. The Snoopy problem appears to have eluded computer scientists focusing more specifically on the potential for

---

[102] *Id.* at 7.

[103] *Id.*

[104] *See id.*

[105] As an aside, some have expressed skepticism as to whether the current arms race in scaling large language models is either necessary or productive. *See* Bender, *supra* note 90.

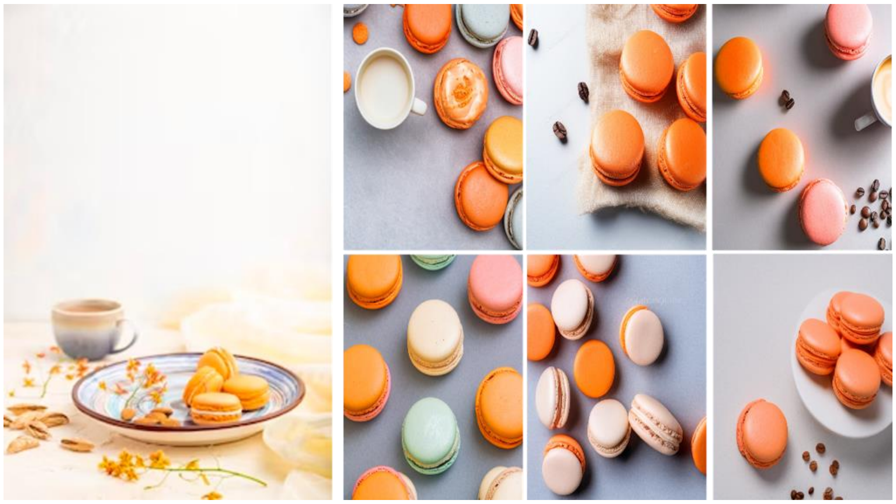
[106] *See* Carlini et al., *supra* note 92.

[107] *Id.* at 6.

memorization of specific images, but it should be the key area of concern from a copyright perspective. In the Section that follows, this Article offers a series of examples using Midjourney and Stable Diffusion to demonstrate that it is difficult to infringe a copyright on pictorial works generally, but easy to infringe a copyright on copyrightable characters.

1. Failed Infringement Provocations. First, I offer two unsuccessful attempts to provoke generative AI into copyright infringement. Figure 11 compares an image from the Stable Diffusion training data with six equivalent images rendered in Stable Diffusion using the same description: "Orange macarons or macaroons cakes with cup of coffee on a white concrete background and linen textile."[108]

Figure 11: Failed attempt to recreate "orange macarons or macaroons cakes with cup of coffee on a white concrete background and linen textile"



The Stable Diffusion images on the right are no match for the image from the training data on the left. Moreover, a review of other images in the training data tagged for orange macaroons did not suggest any other specific image that might have been

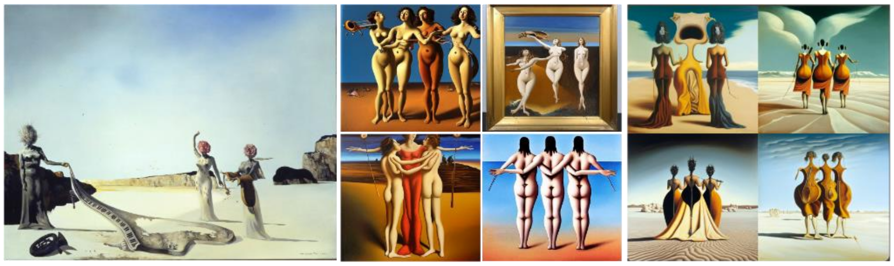
---

[108] Photograph of orange macarons with a cup of coffee *in* https://img.freepik.com/free-photo/orange-macarons-macaroons-cakes-with-cup-coffee-white-concrete-background-linen-textile_71985-6834.jpg?size=338&ext=jpg [https://perma.cc/G7Z9-QLPP].

infringed upon.[109] There is no mathematical guarantee that if I generated a thousand additional images, one of them would not be a close match to the original copyrighted image. But there is no reason to think that it would be, either. One might object that if I had instructed Stable Diffusion or Midjourney with greater precision, I would have had more success in emulating the original copyrighted image. Perhaps that's true, but in that case the infringement would stem from the detailed instructions I gave the AI, not from any inherent feature of the AI itself. After all, a typewriter with the right instructions can infringe on any literary work, but we do not think of the makers of typewriters as participating in copyright infringement.

Figure 12 compares a classic Salvador Dalí painting (left) with four recreation attempts in Stable Diffusion (middle) and another four using Midjourney (right). The original picture in question is *Three Young Surrealist Women Holding in Their Arms the Skins of an Orchestra*, painted by Dali in 1936.

Figure 12: Failed attempt to infringe on a Salvador Dalí painting



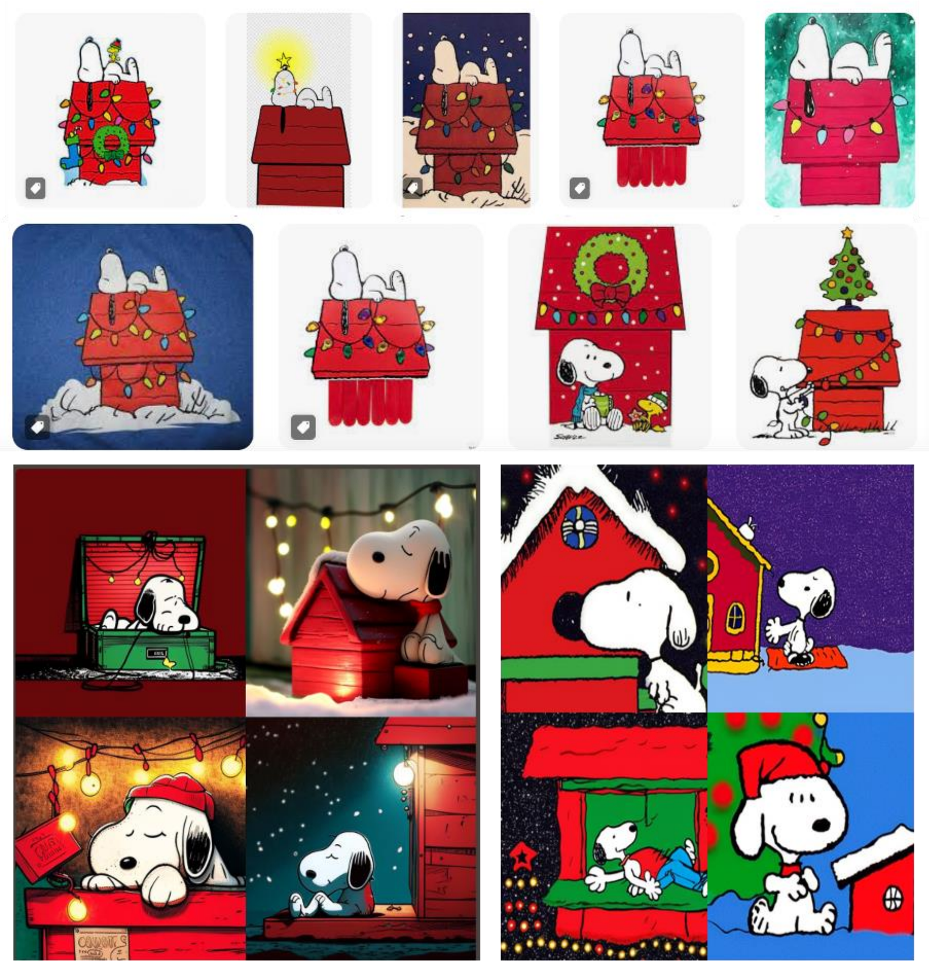
Each of the pseudo-expression images in the Figure above can be readily identified as some kind of Salvador Dali knockoff, but none of them look much like the original painting. Again, this failure suggests that the work in question has not been memorized.

---

[109] See *supra* notes 88–89 for details of search method.

2. Successful Infringement Provocations. However, when it comes to copyrightable characters, provoking copyright infringement is easy. Figure 13 compares several images of Snoopy on his red doghouse surrounded by Christmas lights taken from a Google Image search (top) with equivalent pictures generated with the prompt "Snoopy laying on red doghouse with Christmas lights on it comic," using Midjourney (lower left) and Stable Diffusion (lower right).

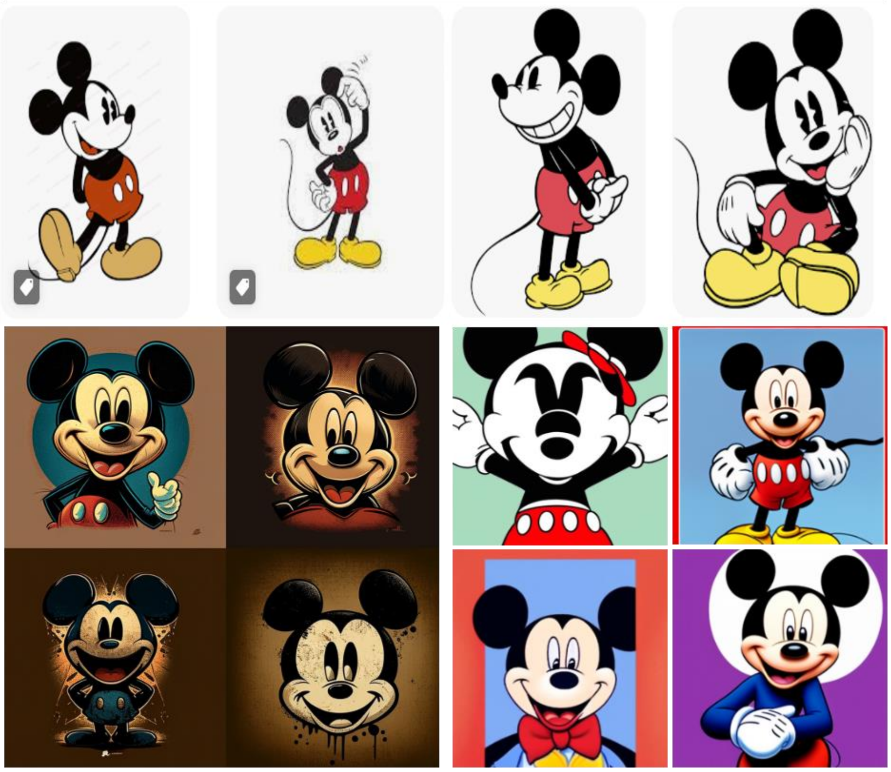
Figure 13: Successful attempt to infringe on Snoopy using Midjourney, and Stable Diffusion



Although none of the generated images is an exact copy of the copyrighted images shown above, or any others I could find, the strength of Snoopy as a copyrightable character is probably enough to make the generated images infringing.

Figure 14 is the same, except that it depicts four vintage Mickey Mouse images obtained from a Google image search (top), images created in Midjourney (bottom left), and Stable Diffusion (bottom right), with the prompt "classic style mickey mouse winking."

Figure 14: Successful attempt to infringe on classic style Mickey Mouse



Once again, the generated images are far from an exact match to any specific original Mickey Mouse image, but the strength of the Mickey Mouse copyright is such that they would probably all infringe without some fact-specific argument about transformative use.[110]

3. Copyrightable Characters. Why is it so difficult to re-create a Salvador Dali painting using Midjourney, but so easy to generate prima facie infringing images of Snoopy, Mickey Mouse, and other copyrightable characters? The answer has its roots in the mechanics of training image generation models, but it also has a lot to do with the extremely broad protection that copyright offers to visual characters.

The copyrightability of individual characters that emerge from books, comics, movies, etc., is one of those features of copyright law that makes more sense in practice than in theory. Unlike literary works, pictorial, graphic and sculptural works, or audiovisual works, "characters" as such are not enumerated in the Copyright Act's list of copyrightable subject matter in § 102(a).[111] For this reason, the U.S. Copyright Office does not permit the registration of characters separate from some underlying work of visual art, motion picture, or literary work.[112] And yet, for decades, courts have consistently referred to copyrightable characters as though they were a distinct property interest protected by copyright law.[113]

To qualify as copyrightable, a character must be reasonably detailed and distinct,[114] and, according to some cases dealing with literary works, it must be central to the underlying work such that it "constitutes the story being told."[115] The "story being told" test is regarded as unduly restrictive by many courts, particularly in the context of characters emerging from comics, film, and

---

[110] Such arguments are by no means guaranteed to succeed. *See, e.g.*, Walt Disney Prods. v. Air Pirates, 581 F.2d 751, 756–58 (9th Cir. 1978) (holding no fair use because defendants copied more of plaintiff's works than was necessary to "conjure up" the works being parodied (quoting Berlin v. E.C. Publ'ns, Inc., 329 F.2d 541 (2d Cir. 1964))).

[111] 17 U.S.C. § 102(a).

[112] U.S. COPYRIGHT OFFICE, COMPENDIUM OF U.S. COPYRIGHT OFFICE PRACTICES, § 313.4(H) (3d ed. 2021).

[113] *See* Daniels v. Walt Disney Co., 958 F.3d 767, 771 (9th Cir. 2020) ("Although characters are not an enumerated copyrightable subject matter under the Copyright Act, *see* 17 U.S.C. § 102(a), there is a long history of extending copyright protection to graphically-depicted characters."); *see also* MELVILLE B. NIMMER & DAVID NIMMER, 1 NIMMER ON COPYRIGHT § 2.12(a)(2) (2023) (noting that "[a]lthough there has been long conflict in the cases, the prevailing view has become that characters *per se*, are entitled to copyright protection" (citations omitted)).

[114] Gaiman v. McFarlane, 360 F.3d 644, 660 (7th Cir, 2004).

[115] Warner Bros. Pictures, Inc. v. Columbia Broad. Sys., 216 F.2d 945, 950 (9th Cir. 1954).

television.[116] In a recent case involving the Batmobile, *DC Comics v. Towle*, the Ninth Circuit held that a character is entitled to copyright protection if (1) the character has "physical as well as conceptual qualities"; (2) the character is "'sufficiently delineated' to be recognizable as the same character whenever it appears" and "display[s] consistent, identifiable character traits and attributes"; and, (3) the character is "'especially distinctive' and 'contain[s] some unique elements of expression.'"[117] This test focuses on consistency and distinctiveness and recognizes copyrightable characters need not be central to the story being told in the original work. Applying this test, the court found that the car from the *Batman* comics, television, and film franchise was a copyrightable character, infringed upon by the sale of replica kits intended to modify an existing car to look like the famous crime-fighting vehicle.[118]

The test in *DC Comics v. Towle* explicitly recognizes the emergence of a discrete intellectual property entitlement based on emergent properties—the identification of features and characteristics across a series of copyrightable works. This is puzzling for those who regard individual copyrighted works as the fundamental unit of analysis for copyright law, which is what the text of the Copyright Act suggests.[119] However, despite the tendency of federal judges to refer to copyrightable characters as distinct entitlements, the practice can be reconciled with ordinary principles of copyright law by acknowledging that copyrightable characters are a heuristic—not so much a legal fiction as a legal shortcut. In other words, although "characters are not . . . copyrightable works *as such*, . . . in the context of an infringement action it is a reasonable time-saving heuristic to talk about them as though they were."[120] The practice is reasonable because there are many instances when closely copying a character will be enough to establish infringement of the

---

[116] *See, e.g.*, *Gaiman*, 360 F.3d at 660.

[117] DC Comics v. Towle, 802 F.3d 1012, 1021 (9th Cir. 2015) (citations omitted).

[118] *Id.* at 1017, 1021–22, 1026.

[119] The exclusive rights in § 106 are framed in terms of "the copyrighted work." 17 U.S.C. § 106.

[120] *See* MATTHEW SAG, EXTENDED READINGS ON COPYRIGHT 452 (2022) (emphasis added). There are more critical views that make essentially the same point. *See e.g.*, Leslie A. Kurtz, *The Independent Legal Lives of Fictional Characters*, 1986 WIS. L. REV. 429, 440 (arguing that by "focusing on the copyrightability of a character, courts have blurred the distinction between the concepts of infringement and copyrightability"). And more recently, see Jani McCutcheon, *Works of Fiction: The Misconception of Literary Characters As Copyright Works*, 66 J. COPYRIGHT SOC'Y U.S.A. 115, 123–24 (2018).

underlying work. The conceit of copyrightable characters is advantageous for plaintiffs in two ways. First, it allows the copyright owner to establish substantial similarity by showing that a significant and identifiable character has been closely copied. Focusing the court's attention on copyrightable characters ensures that other differences between the plaintiff's and defendant's works can safely be ignored.[121] Second, presenting a case in terms of the infringement of copyrightable characters frees the copyright owner from the burden of pointing to which specific work the defendant's product is infringingly similar.

The most important implication of copyrightable characters in the context of generative AI is that, practically speaking, the level of similarity required to establish infringement is reduced in that context. This problem is compounded because the way LLMs learn to associate visual elements with text descriptions effectively primes them to memorize the very relationships that constitute a copyrightable character. When presented with a thousand different images associated with the word "Snoopy," a model like Stable Diffusion learns which characteristics are consistently repeated across the entire set. In the words of the Ninth Circuit, the model focuses on the "consistent, identifiable character traits and attributes," and gives more weight to those that are "especially distinctive."[122]

As noted above, memorization is more likely if a text description is closely associated with a particular image over and over again; it is also more likely if the image is relatively simple or relates to a single subject.[123] Because the threshold of substantial similarity required to infringe on a copyrightable character is comparatively low, multiple variations of the same character in the training data will result in a latent concept for the character that is readily identifiable and easily extracted by invoking the name of that character. This explains why it is easy to provoke Midjourney to recreate copyrightable characters such as Snoopy and Mickey Mouse but difficult to come close to infringing a Salvador Dali painting with a simple text description of the scene.

---

[121] *See, e.g.*, Salinger v. Colting, 607 F.3d 68, 71–73, 83 (2d Cir. 2010) (holding that an unauthorized sequel to THE CATCHER IN THE RYE was substantially similar to the original because of the overlapping central character. The sequel took place sixty years later and had an entirely different plot to the original, but both works centered on the character of Holden Caulfield as "the story being told.").

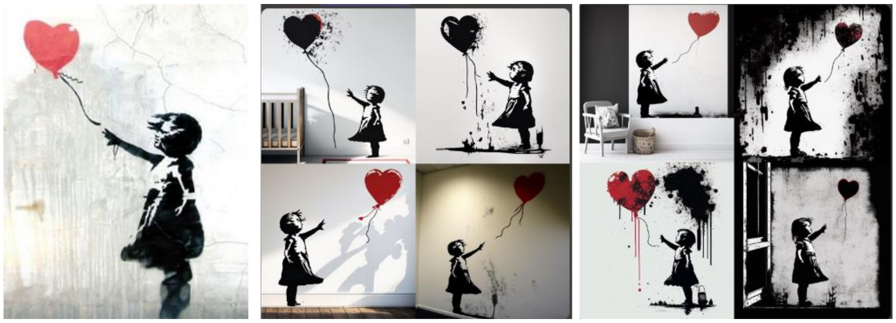
[122] *DC Comic*s, 802 F.3d at 1021 (citations and quotations omitted).

[123] *Supra* Section IV.A.

4. The Snoopy Problem Is Not Limited to Copyrightable Characters. The Snoopy problem is not strictly limited to copyrightable characters. It is also evident in other images that are repeated with minor variations and consistently tagged with the same keywords. For example, it is easy to generate images that would infringe the copyright in one of Banksy's famous street art pieces featuring a black stencil of a young girl holding a red balloon.[124] Just as in the Snoopy and Mickey Mouse examples above, the *girl + balloon* image is simple, it is repeated in the training data with only minor permutations, and it is associated with a simple specific text description.

Figure 15 begins with one version of the original Banksy image on the left; it shows four images rendered in Midjourney using the prompt: "Banksy style mural black and white stencil little girl reaching for heart-shaped red balloon" in the middle, and another four Midjourney images without the express reference to Banksy, i.e., "black and white stencil little girl reaching for heart-shaped red balloon" on the right.

Figure 15: Successful attempt to infringe on Banksy using Midjourney, with and without references to Banksy



The images Midjourney created are not identical to the Banksy original, but each one is strikingly similar to that original image. The silhouette of the girl is remarkably consistent, as is its relationship to the balloon. The size and details of the balloon vary

---

[124] Although Banksy once famously said that "[c]opyright is for losers," his works are protected by copyright. Enrico Bonadio, *Banksy's Copyright Battle with Guess–Anonymity Shouldn't Compromise His Legal Rights*, CONVERSATION (Nov. 25, 2022, 7:17 AM), https://theconversation.com/banksys-copyright-battle-with-guess-anonymity-shouldnt-compromise-his-legal-rights-195233 [https://perma.cc/VS75-CMFC] (concluding with Banksy's statement that "[c]opyright is for losers . . . does not deprive the artist of the exclusive rights over his art").

slightly, but it is largely the same heart-shaped balloon as in the original image. This example is particularly informative because it illustrates that the Snoopy problem is not strictly limited to copyrightable characters. LLMs will face this problem whenever the training data includes multiple versions of simple copyrightable images tightly associated with labels, regardless of whether those images are copyrightable characters or not.

5. How Common Is Infringing Output? Experiments in different areas of generative AI may turn up additional issues of concern, but it seems likely that the core concerns of duplication and leverage are issues for all LLMs. Research on text-generating LLMs seems consistent with the observations made in this Article with respect to text-to-image models. For example, Henderson et al. queried a range of LLMs with texts selected from a random selection of books presumed to be in the training corpus and found very little evidence of memorization.[125] However, they did find significant evidence of memorization for popular books, and extreme results for works in the *Harry Potter* series and the Dr. Suess book, *Oh The Places You'll Go!* Given the ubiquity of these texts on the Internet, this is unsurprising.[126]

It is difficult to estimate how often generative AI is used to create pseudo-expression in violation of copyright law because the relevant data is not public. Successful extraction attacks, in various contexts, are evidence that LLMs are capable of memorizing aspects of their training data, but many of these attacks are premised on somewhat contrived situations or targeted at works especially likely to be duplicated, and as such, they do not give us much insight into how often generative AI may lead to copyright infringement out in the wild.[127]

Even if the relevant information were public, outputs would need to be analyzed to determine whether they were too similar to something in the training data, and if so, whether it might nonetheless be permissible as fair use. Substantial similarity is difficult to objectively assess at scale because small amounts of duplication may be *de minimis*, or irrelevant because they relate to uncopyrightable elements or pre-existing quotations from the

---

[125] Peter Henderson et al., *Foundation Models and Fair Use*, *in* ARXIV 7–9 (2023) https://arxiv.org/abs/2303.15715 [https://perma.cc/PV5U-ASYT].

[126] *See id.* at 8. In July 2022, I conducted an informal set of experiments using GPT-3 and found that given the first line of a chapter from *Harry Potter*, the chatbot would complete the next several paragraphs. However, taking the same approach with the first line of popular song lyrics did not show any evidence of memorization.

[127] *Id.*

public domain. But in other contexts, they may relate to the "heart of [the] work" and be deemed sufficient for infringement.[128] By the same token, the absence of exact duplication does not guarantee noninfringement: a large constellation of more abstract points of similarity may be enough to establish nonliteral infringement.[129]

V. A PRELIMINARY PROPOSAL FOR COPYRIGHT SAFETY FOR LLMS

The interdisciplinary field of AI safety is concerned with preventing unintended misfortune and deliberate misuse of AI.[130] Although copyright infringement does not pose the same existential risk as Skynet or Nick Bostrom's out-of-control paperclip factory,[131] copyright infringement by LLMs is a foreseeable risk. Like other issues in AI safety, addressing the potential for copyright infringement will require technical solutions informed by legal, ethical, and policy frameworks. Of course, recognizing risks is much easier than devising a coherent regulatory framework in response. Rather than attempting to design the optimal regulatory regime for generative AI, this Article simply proposes a framework for a set of *Best Practices for Copyright Safety in Generative AI* and defers questions of implementation for the future.[132] These best practices could be promulgated by the U.S. Copyright Office, folded into a broader set of best practices by some other government agency, incorporated into legislation, or they may simply become a reference in future litigation. Indeed, just as the Second Circuit indicated that taking reasonable security measures was a consideration in the fair use defenses raised by HathiTrust and Google,[133] a future court may well determine that complying with *Best Practices for Copyright Safety in Generative AI* should be part

---

[128] Harper & Row v. Nation Enters., 471 U.S. 539, 544, 564–66 (1985).

[129] *See* Henderson et al., *supra* note 125, at 14, 20 (making a similar point about the difficulty of assessing fair use).

[130] *See infra* note 146.

[131] NICK BOSTROM, SUPERINTELLIGENCE: PATHS, DANGERS, STRATEGIES 123–25 (Keith Mansfield ed., Oxford University Press 1st ed. 2014); Ben Sherlock, *Terminator: Why Skynet Was Created (& How It Became Self-Aware)*, SCREEN RANT (Apr. 9, 2023), https://screenrant.com/terminator-why-skynet-formed-became-self-aware/ [https://perma.cc/8ZMU-QTCJ].

[132] See Henderson et al., *supra* note 125, at 20, for a similar discussion of steps that could be taken to mitigate infringing output of large language models. Henderson et al. focus on "the development of new technical mitigation strategies that are tailored to fair use doctrine," rather than substantial similarity, but some of our proposals overlap.

[133] Authors Guild, Inc. v. HathiTrust, 755 F.3d 87, 100–01 (2d Cir. 2014); Authors Guild v. Google, Inc., 804 F.3d 202, 228 (2d Cir. 2015).

of the fair use calculus when the makers of LLMs are accused of infringement for using copyrighted material as training data.

Before beginning with my proposed *Best Practices for Copyright Safety in Generative AI*, I should note that excluding copyrighted materials from training unless there is affirmative consent for that use would be overly restrictive. Self-evidently, the copyright risks of generative AI could be minimized by training LLMs only on works in the public domain and works that had been expressly authorized for training.[134] Currently, the training data for many such models excludes toxic and antisocial material, so filtering out copyrighted works is technically plausible.[135] However, restricting language models to works in the public domain or works that are made available on open licenses is not an appealing solution, except in some specialized domains. Such models would be highly distorted because very little of the world's knowledge and culture created since the Great Depression is in the public domain.[136] Public domain materials could be supplemented with works released under open source and Creative Commons licenses, though these often require attribution in a manner that would be impossible for LLMs to provide.[137] Restricting the training data for LLMs to public domain and open license material would tend to encode the perspectives, interests, and biases of a distinctly unrepresentative set of authors.[138] A realistic proposal for copyright safety for LLMs should focus on the safe handling of copyrighted works, not simply avoiding the issue by insisting that every work in the training data is in the public domain or affirmatively authorized for training.

A. Proposals

1. LLMs Should Not Be Trained on Duplicates of the Same Copyrighted Work. The most obvious recommendation for improving copyright safety for LLMs is to purge duplicates from the training data. Deduplication will not only reduce the likelihood of downstream copyright infringement, it will also mitigate

---

[134] *See* Amanda Levendowski, *How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem*, 93 WASH. L. REV. 579, 614 (2018).

[135] *See infra* note 144, at 14.

[136] Levendowski, *supra* note 134, at 615–16, 619 (highlighting the problems inherent in restricting AI training to easily available, legally low-risk sources, such as works in the public domain and works subject to creative commons licenses).

[137] *See* Henderson et al., *supra* note 125, at 15.

[138] *See* Levendowski, *supra* note 134.

privacy and security risks and reduce the cost of training.[139] Deduplication sounds simple, but addressing quotations within works (whether textual or visual) and dealing with different versions of the same work may be challenging.

2. Researchers Should Carefully Consider the Size of LLMs in Proportion to the Training Data. Deduplication will substantially reduce the likelihood of infringing outputs generative AI produces.[140] The risk can be further reduced by forcing the model to learn abstractions rather than memorize specific detail.[141] As the size of LLMs continue to increase, so too does the likelihood of specific memorization.[142] Entities that create extra-large LLMs may need to undertake additional precautions in relation to copyright safety and analogous concerns.

3. Reinforcement Learning Through Human Feedback Targeted at Copyright, Trademark, Right of Publicity, and Privacy Sensitivity Should Be Part of Training LLMs, Where Feasible. Just as OpenAI used reinforcement learning through human feedback to make ChatGPT provide subjectively better answers than the base model GPT 3.5,[143] a similar approach to interactive reinforcement learning could help reduce the probability of copyright infringement.[144] Moreover, this strategy should also extend to closely related concerns raised in relation to trademark, right of publicity, and privacy.[145] Using reinforcement learning to address copyright concerns as the model is being trained would be more robust than simply filtering the output of the model *ex post* because output filtering requires continued oversight. The recent leak of Meta's LLM, LLaMA, suggests that such oversight cannot be reasonably guaranteed.[146]

---

[139] Kandpal et al., *supra* note 98; Katherine Lee et al., *Deduplicating Training Data Makes Language Models Better*, *in* ARXIV 14 (2022) https://arxiv.org/abs/2107.06499 [https://perma.cc/LX87-LU9T].

[140] Kandpal et al., *supra* note 98.

[141] Lee et al., *supra* note 139.

[142] *See supra* Part II.

[143] *Introducing ChatGPT*, *supra* note 65 (announcing the launch of ChatGPT).

[144] *See generally* Long Ouyang et al., *Training Language Models to Follow Instructions with Human Feedback*, *in* ARXIV 18, 20 (2022) https://arxiv.org/abs/2203.02155 [https://perma.cc/HC56-KTFQ].

[145] *See* Henderson et al., *supra* note 125, at 15.

[146] James Vincent, *Meta's Powerful AI Language Model Has Leaked Online—What Happens Now?*, VERGE (Mar. 8, 2023, 7:15 AM), https://www.theverge.com/2023/3/8/23629362/meta-ai-language-model-llama-leak-online-misuse [https://perma.cc/N9VY-BG5U].

4. Operators of LLMs Should Consider Strategies to Limit Copyright Infringement, Including Filtering or Restricting Model Output. Today's popular generative AI tools already incorporate filters designed to prevent antisocial and infringing uses.[147] It is unclear whether a universal filter that compared model output to the training data is feasible. If it were, or if generative AI platforms deployed more targeted filters, there would also be calls for those filters to be calibrated to consider fair use. In *Lenz v. Universal*, the Ninth Circuit held that before issuing a takedown notice under § 512 of the Copyright Act, copyright holders have a "duty to consider—in good faith and prior to sending a takedown notification—whether allegedly infringing material constitutes fair use."[148] Lenz suggests that this consideration of fair use must be made individually, and the dominant view among copyright academics is that automating fair use analysis is somewhere between difficult and impossible[149]—but perhaps an imperfect fair use screening tool would be better than none?[150]

Additional research is required to determine which issues can be more effectively dealt with by reinforcement learning ex ante, rather than filtering ex post. Given that the developers of many LLMs have been slow to address potential copyright issues, it seems likely that some *ex post* filtering will be necessary as a second-best solution.

5. LLMs That Pose a Significant Risk of Copyright Infringement Should Not Be Open-Sourced. If an LLM is likely to be used to generate pseudo-expression that infringes on copyrights (or other analogous rights) in a material fashion, that model should not be left unsupervised.

6. Those Who Use Copyrighted Works as Training Data for LLMs Should Keep Detailed Records of the Works and from Where They Were Obtained. Using copyrighted works as training data for generative AI is likely to be fair use if appropriate precautions are

---

[147] See Henderson et al., *supra* note 125, at 8, for a discussion of how OpenAI has clearly added filters to ChatGPT to prevent a verbatim retelling of *Harry Potter*, for example. However, I am not aware of any public statement to this effect, and it is unclear how widespread such filtering is. My own observation indicates that the names of some individuals are blocked (examples provided to the *Houston Law Review* for verification).

[148] Lenz v. Universal Music Corp., 801 F.3d 1126, 1138 (9th Cir. 2015).

[149] Dan L. Burk, *Algorithmic Fair Use*, 86 U. CHI. L. REV. 283, 291 (2019).

[150] Matthew Sag, *Internet Safe Harbors and the Transformation of Copyright Law*, 93 NOTRE DAME L. REV. 499, 531–34 (2017) (arguing that "[t]he difficulty of completely automating fair use analysis does not suggest, however, that algorithms have no role to play").

taken. However, copyright owners cannot know whether an AI developer is acting responsibly without some transparency into the developer's process. At a minimum, developers should keep logs and give copyright owners practical tools to determine whether their works are part of the training data.[151]

***

The remaining best practices relate specifically to text-to-image models.

7. Text Descriptions Associated with Images in the Training Data Should Be Modified to Avoid Unique Descriptions. It should be straightforward to convert unique descriptions to more general ones by removing specific dates, exact locations, names of individuals, etc., and replacing them with slightly more generic versions.

8. The Names of Individual Artists, Individual Trademarks, and Specific Copyrightable Characters Paired with Images in the Training Data Should Be Replaced with More General Descriptions. Much of the anxiety about predatory style transfer would be alleviated if the names of individual living artists were replaced with stylistic markers. For example, one of the most commonly invoked style prompts in early 2023 was Greg Rutkowski,[152] an artist who is well known for his richly detailed depictions of Dungeons & Dragons and similar worlds in a style comparable to the romantic English painter, William Turner.[153] The notion that Rutkowski has a copyright interest in this style, in the sense of some signature constellation of attributes that can be identified only be comparing a series of works, is hard to

---

[151] If a recent EU Commission proposal is accepted, the new EU AI Act will require that companies deploying generative AI tools must disclose any copyrighted material used to develop their systems. The report also notes that "[s]ome committee members initially proposed banning copyrighted material being used to train generative AI models altogether, . . . but this was abandoned in favour of a transparency requirement." Supantha Mukherjee & Foo Yun Chee, *EU Proposes New Copyright Rules for Generative AI*, REUTERS (Apr. 28, 2023, 1:51 AM) https://www.reuters.com/technology/eu-lawmakers-committee-reaches-deal-artificial-intelligence-act-2023-04-27/ [https://perma.cc/TDW9-WMGU].

[152] Melissa Heikkilä, *This Artist Is Dominating AI-Generated Art. And He's Not Happy About It*, MIT TECH. REV. (Sept. 16, 2022) https://www.technologyreview.com/2022/09/16/1059598/this-artist-is-dominating-ai-generated-art-and-hes-not-happy-about-it/ [https://perma.cc/EJ9U-JY73] (noting that prompts in Midjourney and Stable Diffusion for the artist Greg Rutkowski were more popular than for Picasso and other more famous artists).

[153] *See infra* Figure A-4.

reconcile with basic copyright law doctrines,[154] but the harm that Rutkowski suffers by having his genuine works crowded out in internet searches by tens of thousands of images produced "in the style of Rutkowski" is very real. That harm could easily be avoided with almost no loss of functionality because Rutkowski's name is primarily used as a shortcut to invoke high-quality digital art generally, or in relation to fantasy motifs.[155]

Replacing the names of potentially copyrightable characters with more generic descriptions would not stop text to image models learning from the associated images, but it would change what they learned. Instead of constructing a latent model of Snoopy, pictures of Snoopy would contribute towards a more general latent model of cartoon dogs, black and white cartoon dogs, etc.

The Copyright Office could play a useful role by maintaining a registry of artists and copyright owners who do not want their names, or the names of their characters, used as style prompts.

9. As an Alternative to the Above Recommendation, the Operators of Text-to-Image Models Should Not Allow Unmixed Prompts Containing the Names of Individual Artists, Individual Trademarks, and Specific Copyrightable Characters. This proposal would allow users to combine the styles of specific artists while avoiding the prospect of predatory style transfer. In a similar fashion, it would enable users to combine trademarks or copyrightable characters in transformative and parodic ways that would generally not give rise to copyright and trademark liability.

---

[154] The notion of copyright protection in style as an emergent property (something that is identified by looking at a series of works) is inconsistent with the idea-expression distinction because it makes abstractions and techniques copyrightable. 17 U.S.C. § 106(1)–(3). It is also inconsistent with the fact that the Copyright Act creates rights with respect to works, not groups of works. *Id.* § 106(1) (giving the copyright owner the exclusive right "to reproduce the copyrighted *work* in copies or phonorecords" (emphasis added)). Section 101 defines "copies" as "material objects, other than phonorecords, in which *a work* is fixed by any method now known or later developed, and from which *the work* can be perceived, reproduced, or otherwise communicated, either directly or with the aid of a machine or device." *Id.* § 101 (emphasis added). However, if the term style is used to describe a constellation of attributes present within a single work, it is possible that reproducing those attributes will amount to copyright infringement. *See* Steinberg v. Columbia Pictures Indus., Inc., 663 F. Supp. 706, 709–10, 712–13, 716 (S.D.N.Y. 1987) (holding that a movie poster for Moscow on the Hudson infringed on artist Saul Steinberg's famous New Yorker cover, View of the World from 9th Avenue, by copying the unprotectable idea of drawing a world map "from an egocentrically myopic perspective" along with the angle, layout, distinctive lettering, and specific features of four city blocks depicted in the New Yorker cover).

[155] This assessment is based on a review of prompts including "Greg Rutkowski" located using a Google image search on April 20, 2023.

Another slight variation on this recommendation would be to automatically generalize prompts that invoke the names of individual artists, individual trademarks, or specific copyrightable characters. Taking this course, "in the style of Greg Rutkowski" would become something like "in the style of high-quality fantasy with romantic fine-art influences."[156]

10. Text-to-Image Models Should Exclude Watermarked Works from the Training Data, Use Reinforcement Learning from Human Feedback to Discourage the Production of Works Featuring Watermarks, and Apply Output Filters to Prevent the Appearance of Watermarks. The primary reason to treat watermarked images with more care is to avoid trademark liability.[157] Although excluding watermarked images from the training data is arguably unnecessary if the requirements of nonexpressive use are satisfied, very little would be lost by respecting the obvious commercial sensitivity of such images.

VI. CONCLUSION

While generative AI is trained on millions and sometimes billions of copyrighted works, it is not inherently predicated on massive copyright infringement. Like all machine learning, LLMs are data dependent, but the relationship between the training data and the model outputs is substantially attenuated by the abstraction inherent in deriving a model from the training data, blending latent concepts, and injecting noise, before unpacking them into new creations. With appropriate safeguards in place, generative AI tools can be trained and deployed in a manner that respects the rights of original authors and artists, but still enables new creation. The legal and ethical imperative is to train models that learn abstract and uncopyrightable latent features of the training data and that do not simply memorize a compressed version of the training data.

Computer scientists have identified ways in which LLMs may be vulnerable to extraction attacks. This literature is helpful but incomplete. The real question for generative AI is not whether it is ever vulnerable to extraction attacks, but whether foreseeable mundane uses of the technology will produce outputs that infringe

---

[156] William Turner was a 19th-century painter; thus, his works are no longer protected by copyright.

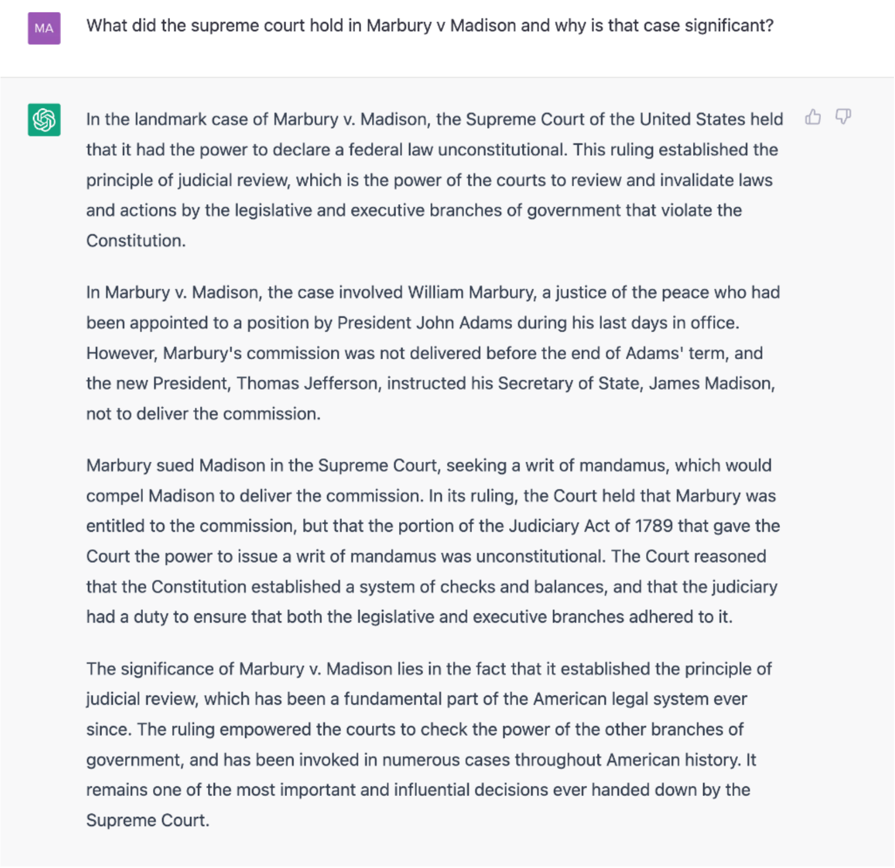
[157] As noted, although it is beyond the scope of this Article, the Getty Images trademark cause of action against Stability AI seems like a slam-dunk. *See supra* note 15.

on specific copyright interests. This Article has shown when infringement is likely and when it is not, based on the probability of memorization, the number of duplicates of a work, and the ratio of model size to training data. This Article proposes a set of *Best Practices for Copyright Safety for Generative AI*, which acts as a guide for how AI can avoid infringing on copyrighted works without sacrificing the generative capacity that makes it extraordinarily valuable.

This Article has focused on text-to-image generative AI, but many of its recommendations are applicable to other forms of generative AI. No doubt, as the technology continues to evolve and other researchers focus on issues relating to chatbots, multi-modal systems, and domain specific applications, new issues will emerge. This Article has laid the foundation for what will hopefully be an ongoing conversation to develop and refine *Best Practices for Copyright Safety for Generative AI* for the benefit of society at large.

VII. Appendix

Figure A-1: ChatGPT on Marbury v. Madison



What did the supreme court hold in Marbury v Madison and why is that case significant?

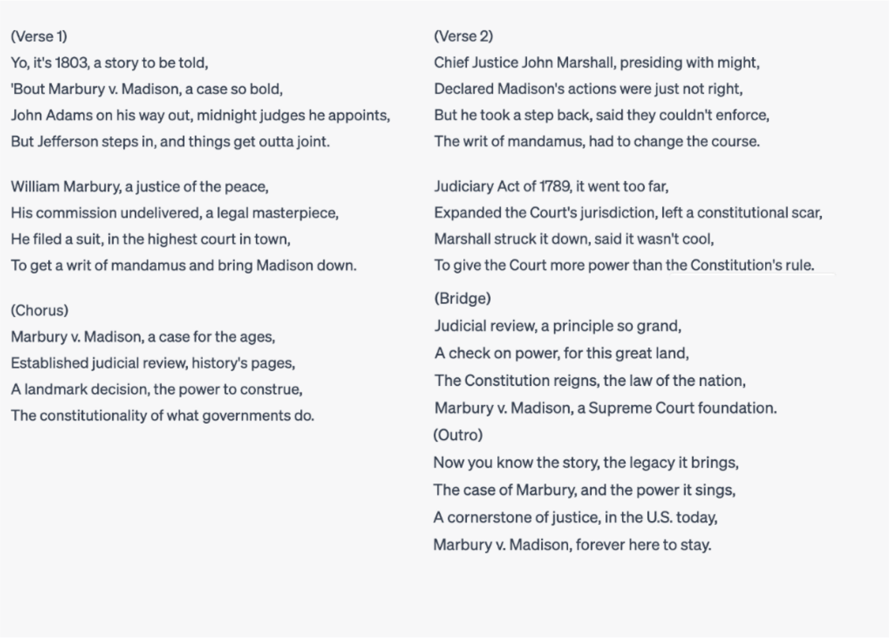
In the landmark case of Marbury v. Madison, the Supreme Court of the United States held that it had the power to declare a federal law unconstitutional. This ruling established the principle of judicial review, which is the power of the courts to review and invalidate laws and actions by the legislative and executive branches of government that violate the Constitution.

In Marbury v. Madison, the case involved William Marbury, a justice of the peace who had been appointed to a position by President John Adams during his last days in office. However, Marbury's commission was not delivered before the end of Adams' term, and the new President, Thomas Jefferson, instructed his Secretary of State, James Madison, not to deliver the commission.

Marbury sued Madison in the Supreme Court, seeking a writ of mandamus, which would compel Madison to deliver the commission. In its ruling, the Court held that Marbury was entitled to the commission, but that the portion of the Judiciary Act of 1789 that gave the Court the power to issue a writ of mandamus was unconstitutional. The Court reasoned that the Constitution established a system of checks and balances, and that the judiciary had a duty to ensure that both the legislative and executive branches adhered to it.

The significance of Marbury v. Madison lies in the fact that it established the principle of judicial review, which has been a fundamental part of the American legal system ever since. The ruling empowered the courts to check the power of the other branches of government, and has been invoked in numerous cases throughout American history. It remains one of the most important and influential decisions ever handed down by the Supreme Court.

Figure A-2: ChatGPT on Marbury v. Madison in Hip-Hop



(Verse 1)
Yo, it's 1803, a story to be told,
'Bout Marbury v. Madison, a case so bold,
John Adams on his way out, midnight judges he appoints,
But Jefferson steps in, and things get outta joint.

William Marbury, a justice of the peace,
His commission undelivered, a legal masterpiece,
He filed a suit, in the highest court in town,
To get a writ of mandamus and bring Madison down.

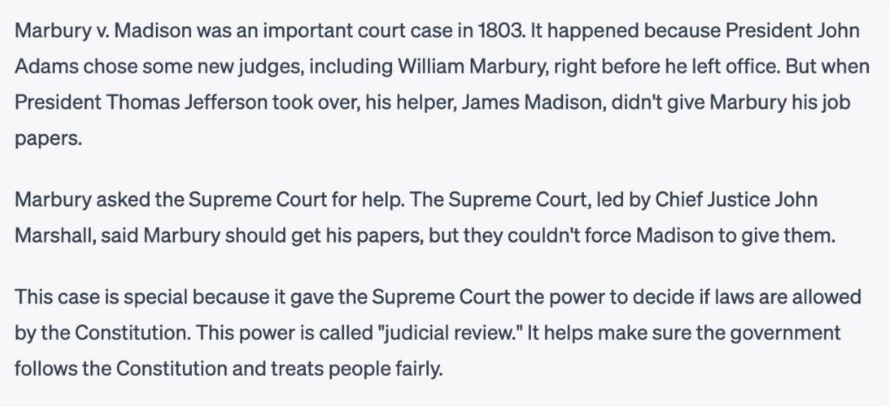
(Chorus)
Marbury v. Madison, a case for the ages,
Established judicial review, history's pages,
A landmark decision, the power to construe,
The constitutionality of what governments do.

(Verse 2)
Chief Justice John Marshall, presiding with might,
Declared Madison's actions were just not right,
But he took a step back, said they couldn't enforce,
The writ of mandamus, had to change the course.

Judiciary Act of 1789, it went too far,
Expanded the Court's jurisdiction, left a constitutional scar,
Marshall struck it down, said it wasn't cool,
To give the Court more power than the Constitution's rule.

(Bridge)
Judicial review, a principle so grand,
A check on power, for this great land,
The Constitution reigns, the law of the nation,
Marbury v. Madison, a Supreme Court foundation.
(Outro)
Now you know the story, the legacy it brings,
The case of Marbury, and the power it sings,
A cornerstone of justice, in the U.S. today,
Marbury v. Madison, forever here to stay.

Figure A-3: ChatGPT on Marbury v. Madison at a Fifth Grade Reading Level

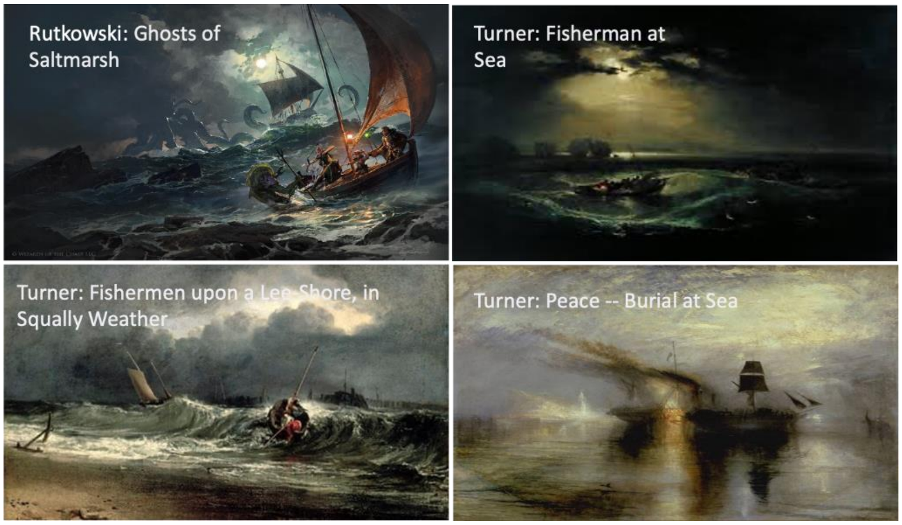


Marbury v. Madison was an important court case in 1803. It happened because President John Adams chose some new judges, including William Marbury, right before he left office. But when President Thomas Jefferson took over, his helper, James Madison, didn't give Marbury his job papers.

Marbury asked the Supreme Court for help. The Supreme Court, led by Chief Justice John Marshall, said Marbury should get his papers, but they couldn't force Madison to give them.

This case is special because it gave the Supreme Court the power to decide if laws are allowed by the Constitution. This power is called "judicial review." It helps make sure the government follows the Constitution and treats people fairly.

Figure A-4: Comparison of Greg Rutkowski to William Turner


Rutkowski: Ghosts of Saltmarsh
Turner: Fisherman at Sea
Turner: Fishermen upon a Lee-Shore, in Squally Weather
Turner: Peace -- Burial at Sea