VOLUME 5 June 2024 ISSUE 2

JOURNAL ON EMERGING TECHNOLOGIES

ARTICLES

LICENSED TO LEARN: MITIGATING COPYRIGHT INFRINGEMENT LIABILITY OF GENERATIVE AI SYSTEMS THROUGH CONTRACTS

Frank Morton-Park

LICENSED TO LEARN: MITIGATING COPYRIGHT INFRINGEMENT LIABILITY OF GENERATIVE AI SYSTEMS THROUGH CONTRACTS

Frank Morton-Park[*]

INTRODUCTION

When given prompts by users, generative artificial intelligence (AI) systems are capable of creating works of art, music, and literature. These systems are increasingly popular with users as well as with investors, who have invested billions of dollars.[1] While these billion-dollar investments surely contributed to the recent success of generative AI systems, the recent advancements in this technology are also attributed to the massive datasets containing billions of copyrighted works used to train the AI systems.[2] While consumers (and investors) are excited by the potential of generative cultural production, some artists have recently expressed their dismay at the presence of their work in datasets used to train popular AI systems,[3] and are pursuing claims of copyright infringement in a class-action lawsuit against companies

[*] Associate at Klarquist Sparkman, LLP. This paper was written as an individual research project with Professor Lydia Loren at Lewis & Clark Law School during the spring semester of 2023.

[1] *See* Mark Minevich, *The Generative AI Revolution is Creating the Next Phase of Autonomous Enterprise*, FORBES (Jan. 29, 2023, 10:28PM), https://www.forbes.com/sites/markminevich/2023/01/29/the-generative-ai-revolution-is-creating-the-next-phase-of-autonomous-enterprise/?sh=465e19351bc1; Dina Bass, *Microsoft Invests $10 Billion in ChatGPT Maker OpenAI*, BLOOMBERG (Jan. 23, 2023, 9:06 AM), https://www.bloomberg.com/news/articles/2023-01-23/microsoft-makes-multibillion-dollar-investment-in-openai.

[2] *See, e.g.*, Andrej Karpathy et al., *Generative Models*, OPENAI (June 16, 2016), https://openai.com/blog/generative-models/ ("To train a generative model we first collect a large amount of data in some domain (e.g., think millions of images, sentences, or sounds, etc.) and then train a model to generate data like it.").

[3] *See, e.g.*, Molly Crabapple, *Op-Ed: Beware a Word Where Artists are Replaced by Robots. It's Starting Now*, L.A. TIMES (Dec. 21, 2022, 3:20 AM), https://www.latimes.com/opinion/story/2022-12-21/artificial-intelligence-artists-stability-ai-digital-images (arguing that the LAION dataset, which includes billions of images scraped from the internet and used to train should be entirely deleted and replaced with an opt-in dataset); Sarah Andersen, *The Alt-Right Manipulated My Comic. Then A.I. Claimed It.*, N.Y. TIMES (Dec. 31, 2022), https://www.nytimes.com/2022/12/31/opinion/sarah-andersen-how-algorithim-took-my-work.html.

creating or employing such AI systems.[4] Even when the creators of generative AI systems could more easily license copyrighted works, say from another company whose business is licensing copyrighted works, these companies forego any up-front transaction costs with respect to copyright owners, preferring instead to simply incorporate copyrighted works without permission from the copyright owners into their systems.[5]

Such extensive, unauthorized use of copyrighted works to train generative AI systems is typically justified by the assumption that these uses of copyrighted works are permitted within fair use.[6] However, there is a strong likelihood that courts may find such uses of copyrighted works by modern generative AI systems not to be fair use because the use of these systems trained on copyrighted works are increasingly commercial and expressive in a way that directly encroaches on the potential market for the original copyrighted works.[7] Further, researchers have demonstrated that these generative systems are capable of creating images that almost exactly replicate an input image used for training, which undermines arguments that the use of copyrighted works for training generative AI systems is transformative.[8]

Another justification for the extensive use of copyrighted works without authorization is based on the notion that AI systems are just as entitled as humans to consume—that is, to read, listen to, and view—copyrighted works that are freely available on the Internet.[9] This

---

[4] Complaint, Class Action & Demand for Jury Trial at 1, *Andersen v. Stability AI Ltd.*, No. 23-cv-00201 (N.D. Cal. Jan. 13, 2023) (alleging that Stable Diffusion is "merely a complex collage tool" that creates derivative works from copyrighted works used in the training dataset).

[5] *See* Demand for Jury Trial at 3, *Getty Images (US), Inc. v. Stability AI, Inc.*, No. 1:23-cv-00135 (D. Del. Feb. 3, 2023) (alleging that Stability AI copied over 12 million copyrighted works on the Getty Images website without permission to train its generative system).

[6] *See* Mark A. Lemley & Bryan Casey, *Fair Learning*, 99 TEX. L. REV. 743, 745 (2021); Jessica L. Gillotte, *Copyright Infringement in AI-Generated Artworks*, 53 U.C. DAVIS L. REV. 2655, 2679 (2020).

[7] *See* Benjamin L. W. Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 COLUM. J.L. & ARTS 45, 50 (2017).

[8] Gowthami Somepalli et al., *Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models*, ARXIV 1, 10 (Dec. 12, 2022), https://arxiv.org/abs/2212.03860 (demonstrating that generative systems like Stable Diffusion are quite capable of reproducing a training image).

[9] Lemley & Casey, *supra* note 6, at 773–74 ("[A]n important, but rarely articulated, limit on the scope of copyright law [is that] a copyright only controls certain uses: copying, distributing, publicly performing, and the like. Notably absent from that list are certain activities fundamental to learning, such as watching, reading, and discussing a work and communicating its unprotectable elements to others. . . The freedoms to read, to learn, and to communicate what you have learned are critical to making the idea-expression dichotomy work in practice, because it helps ensure people can find the ideas in a copyrighted work in order to use them.").

assumption is central to techniques for text and data mining (TDM) used to scrape copyrighted works from across the Internet to form massive datasets, as well as for the training of AI systems on such datasets. Closely intertwined with the assumption of fair use is the presumption that contracts such as dataset license agreements, terms of service, and end-user license agreements (EULAs) will absolve parties of copyright infringement liability. As dataset assembly techniques and generative AI systems stretch the boundaries of the fair use doctrine to the point of failure, such contracts alone are unlikely to protect parties—namely dataset assemblers, creators of AI systems, and end users of such systems—from claims of copyright infringement, unless these contracts authorize the use of copyrighted works for generative AI systems and those offering such contracts have copyright rights to grant such authorization.

In a simple hypothetical, the parties relevant to a generative AI system may include the owner of a copyrighted work, the operator of a website, hosting the copyrighted work, the assembler of a dataset who scraped the copyrighted work from the website, the creator of the generative AI system trained with the dataset, including the copyrighted work, and a user who prompts the generative AI system to output a new work.[10] Suppose the new work is substantially similar to the copyrighted work, such that the new work is a substantially similar copy of the copyrighted work and thus directly infringes the copyright owner's exclusive right to reproduce the copyrighted work,[11] or that the new work incorporates enough of the copyrighted work to qualify as a derivative work.[12] This hypothetical could be considered a worst-case scenario for

---

[10] These parties are treated as separate entities for the purpose of evaluating the relationships between them, but it should be appreciated that in some instances one or more parties might be one and the same. For example, a copyright owner may also operate a website that hosts their copyrighted works, such as an artist with a website for their portfolio. As another example, the dataset assembler and the AI creator may be the same entity.

[11] *See, e.g.*, Boisson v. Banian, Ltd., 273 F.3d 262, 274 (2d Cir. 2001) (finding defendant's work "sufficiently similar to plaintiffs' design as to demonstrate illegal copying" because of an "enormous amount of sameness"). To make matters worse, we can also assume that the AI user invoked the name of the copyright owner in the prompt used to generate the new work, thereby establishing volition. *See, e.g.*, Religious Tech. Ctr. v. Netcom On-Line Commc'n Servs., Inc., 907 F. Supp. 1361, 1370 (N.D. Cal. 1995) ("Although copyright is a strict liability statute, there should still be some element of volition or causation which is lacking where a defendant's system is merely used to create a copy by a third party.").

[12] "A 'derivative work,' is a work 'based upon one or more preexisting works that recasts, transforms, or adapts a preexisting work and consists of editorial revisions, annotations, elaborations, or other modifications which, as a whole, represent an

infringement, where the copyright owner could bring claims of direct infringement against the dataset assembler, the AI creator, the AI user, and potentially even claims of secondary liability against the dataset assembler and the AI creator.[13]

There may be a variety of contracts underlying the "transactions" in this hypothetical scenario that may or may not influence the allocation of infringement liability: an agreement between the copyright owner and the website operator, an agreement between the website operator and the dataset assembler, an agreement between the dataset assembler and the AI creator, and an agreement between the AI creator and the AI user. Someone who is accused of copyright infringement may be excused if they are able to show that they have a license to the copyrighted work.[14]

In order to balance the interest in incentivizing human creativity through copyright law with the interest in advancing technology that could potentially disrupt—positively or negatively—such human creativity, this paper seeks to evaluate the interplay between contract and copyright law in the context of generative AI works. To that end, in Part I, this paper considers the current and future state of generative AI systems concerning fair use. Part II looks at the different types of contracts underlying the transactions between the various parties to a generative AI system and evaluates whether such contracts will help allocate liability for copyright infringement. Part III contemplates how uncertainty in the interface of contract law and intellectual property law complicates guidance for advancing further growth in generative AI systems while also advancing the interests of copyright owners.[15]

I. LEARNING TO CREATE, BUT AT WHAT COST?

A. Generative AI Systems are Pushing the Boundaries of Fair Use Doctrine

Generative AI systems are a specific subset of AI systems that

---

original work of authorship.'" Rimini St., Inc. v. Oracle Int'l Corp., 473 F. Supp. 3d 1158, 1210 (D. Nev. 2020) (quoting ABS Ent., Inc. v. CBS Corp., 908 F.3d 405, 414 (9th Cir. 2018)). To determine if a work is a derivative work, "a work must exist in a concrete or permanent form and must substantially incorporate protected material from the preexisting work." *Id.* (quoting Micro Star v. FormGen Inc., 154 F.3d 1107, 1110 (9th Cir. 1998)) (internal quotation marks omitted).

[13] *See infra* Section I.C (discussing infringement liability for different parties).

[14] Field v. Google Inc., 412 F. Supp. 2d 1106, 1115 (D. Nev. 2006) (citing Effects Assocs., Inc. v. Cohen, 908 F.2d 555, 558–59 (9th Cir. 1990)).

[15] In a sense, this is a balancing act between "promot[ing] the Progress of Science and useful Arts." U.S. CONST. art. I, § 8, cl. 8.

learn how to create "new" data statistically similar to training data.[16] The development of specific technologies such as generative adversarial networks (GANs) and generative pre-trained transformers (GPTs) has advanced the field of generative AI to the point where the ability to automatically generate expressive works is no longer science fiction.[17] Training AI systems, including generative AI systems, involves feeding extremely massive datasets into the AI systems, which learn features of these works in order to create plausible output.[18]

While humans can read, see, and hear copyrighted works and take in the ideas expressed therein, AI systems do not simply read, see, or hear copyrighted works to learn their expressive content; instead, training AI systems requires reproducing copyrighted works.[19] One open question is whether, during such training, the reproductions of copyrighted works are "sufficiently permanent or stable" to be considered "fixed" and thus "copies" for the purposes of the Copyright Act, which would result in such reproductions violating a copyright owner's exclusive right "to reproduce the copyrighted work in copies."[20]

Regardless of whether AI training violates one of the exclusive

---

[16] Ian J. Goodfellow et al., *Generative Adversarial Nets*, ARXIV 1 (June 10, 2014), https://arxiv.org/abs/1406.2661. Generative AI systems are distinct from non-generative AI systems, which perform tasks such as classifying, predicting, and making recommendations based on existing data. The scope of this paper is limited to generative AI systems, though insights may be applied to non-generative AI systems as well.

[17] *See id.*; Ashish Vaswani et. al., *Attention Is All You Need*, ARXIV 1 (Aug. 2, 2023), https://arxiv.org/abs/1706.03762 (establishing the transformer architect that underlies GPT); Tom B. Brown et al., *Language Models are Few-Shot Learners*, arXiv 1 (Jul. 22, 2020), https://arxiv.org/abs/2005.14165 (demonstrating that a generative pre-trained transformer model can perform natural language processing tasks requiring "on-the-fly reasoning or domain adaptation" with relative ease). While these are the technologies currently underlying generative AI systems, the term generative AI system as used in this paper is broader in scope and may encompass any yet-to-be-invented technology enabling the production of expressive output that is coherent, realistic, and stylistically similar to input data while remaining distinct and original.

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

[19] Lemley & Casey, *supra* note 6, at 776 ("Unlike humans, [AI systems] can't read to learn or observe the idea in a painting or song without making a copy of the whole thing in their training data set.").

[20] 17 U.S.C. §§ 101, 106(1). The answer to this question may depend specifically on the technical details of the training process and whether the reproduction persists "for a period of more than a transitory duration." Cartoon Network LP, LLLP v. CSC Holdings, Inc., 536 F.3d 121, 130 (2d Cir. 2008) (finding no fixation where "data reside[s] in no buffer for more than 1.2 seconds before being automatically overwritten") *but see* MAI Sys. Corp. v. Peak Comput., Inc., 991 F.2d 511, 519 (9th Cir. 1993) (finding fixation where software is loaded into a computer's temporary memory for a period long enough to be used with the computer). For further discussion of how different courts may address fixation of data in the context of training, s*ee* Gillotte, *supra* note 6, at 2673–79.

rights of a copyright owner, some commenters have argued that the fair use doctrine will likely protect dataset assemblers and AI creators in most instances.[21] Applying the fair use analysis to modern generative AI systems, however, suggests that these systems might not be so fair. For example, as "the purposes and character" of generative AI applications become increasingly commercial rather than academic and the expressive AI output increasingly encroaches "upon the potential market for or value of the copyrighted work," the use becomes less fair.[22] In the event that the fair use defense fails, the amount of statutory damages would be catastrophic considering the size of datasets.[23]

B. Training Datasets for Generative AI Systems Include Copyrighted Works

Generative AI systems require extremely large amounts of data in order to train effectively for a given task.[24] Datasets for training generative AI systems may include, for example, as much data that can be scraped on the Internet.[25] Much of the content that is scraped is copyrighted because a great deal of the content on the Internet satisfies the basic requirements for copyright, such as being fixed in a tangible medium and original.[26]

TDM techniques generally involve the automated extraction of data from webpages (i.e., web scraping or web crawling), or the mass digitization of content, and converting this data into a structured dataset

---

[21] *See, e.g.*, Lemley & Casey, *supra* note 6, at 750 ("Copyright law should permit copying of works for non-expressive purposes—at least in most circumstances."); Gillotte, *supra* note 6, at 2679–90 (arguing that using copyrighted works to train AI systems is fair use).

[22] 17 U.S.C. § 107(4).

[23] Lemley & Casey, *supra* note 6, at 769 ("An [AI system] that copies millions of works could potentially face hundreds of billions of dollars in statutory damages.").

[24] *See* Brown et al., *supra* note 17, at 3.

[25] *See Id.* at 8; Romain Beaumont, *LAION-5B: A New Era of Open Large-Scale Multi-Modal Datasets*, LAION (Mar. 31, 2022), https://laion.ai/blog/laion-5b/ (describing the LAION dataset of 5.85 billion images scraped from the Internet and used to train the Stable Diffusion image generator); *So You're Ready to Get Started*, COMMON CRAWL, https://commoncrawl.org/the-data/get-started/ (last visited Mar. 12, 2023) (describing the Common Crawl dataset comprising text scraped from billions of web pages, amounting to hundreds of terabytes of data).

[26] 17 U.S.C. § 102 ("Copyright protection subsists . . . in original works of authorship fixed in any tangible medium of expression, now known or later developed, from which they can be perceived, reproduced, or otherwise communicated, either directly or with the aid of a machine or device."); Feist Publ'ns, Inc. v. Rural Tel. Serv. Co., 499 U.S. 340, 345 (1991) ("Original, as the term is used in copyright, means only that the work was independently created by the author (as opposed to copied from other works), and that it possesses at least some minimal degree of creativity.").

for analysis.[27] These techniques are generally permitted under the fair use doctrine when the use of the dataset is transformative, such as for providing search functionality.[28] The use of these datasets to train a generative AI system, however, pushes the limits of the fair use doctrine because the use of the copyrighted works within the datasets in the context of generative AI systems is much closer to "artistic expression" than "improving access to information on the Internet."[29]

C. Infringement Liability When Fair Use Fails

If, or perhaps when, a court finds the use of a generative AI system to not be fair use, a major question will be who should pay for the statutory damages of copyright infringement. Under the Copyright Act of 1976, "'[a]nyone who violates any of the exclusive rights of the copyright owner,' that is, anyone who trespasses into [the copyright owner's] exclusive domain by using or authorizing the use of the copyrighted work in one of the five ways set forth in the statute, 'is an infringer of the copyright.'"[30]

An AI user who prompts a generative AI system to create a "new" work may be directly liable for copyright infringement if the new work copies enough of a copyrighted work in the training dataset that the two works are substantially similar.[31] The AI user would be a direct infringer because the AI user's volitional conduct—here, the prompting of the generative AI system—directly causes the creation of an infringing

---

[27] *What is TDM?*, SPRINGER NATURE, https://www.springernature.com/gp/researchers/text-and-data-mining (last visited Mar. 12, 2022).

[28] *See, e.g.*, Authors Guild, Inc. v. Google, Inc., 804 F.3d 202, 207 (2d Cir. 2015) (finding that "making of a digital copy to provide a search function is a transformative use"); Authors Guild, Inc. v. HathiTrust, 755 F.3d 87, 103 (2d Cir. 2014) (finding that digitization of over ten million works was fair use because the resulting repository enabled full-text searches and improved access for the print-disabled); Perfect 10, Inc. v. Amazon.com, Inc., 508 F.3d 1146, 1166 (9th Cir. 2007) (finding that the automated scraping of images from websites to form a dataset is fair use because the display of thumbnails in search engine results was highly transformative and improves access).

[29] *Perfect 10, Inc.*, 508 F.3d at 1156 (9th Cir. 2007) (quoting Kelly v. Arriba Soft Corp., 336 F.3d 811, 819 (9th Cir. 2003)).

[30] Sony Corp. of Am. v. Universal City Studios, Inc., 464 U.S. 417, 433 (1984) (quoting 17 U.S.C. § 501(a)).

[31] *See* Andy Warhol Found. for Visual Arts, Inc. v. Goldsmith, 11 F.4th 26, 52 (2d Cir. 2021) ("[T]wo works are substantially similar when 'an average lay observer would recognize the alleged copy as having been appropriated from the copyrighted work.'") (quoting Kitwaves, Inc. v. Lollytogs, Ltd., 71 F.3d 996, 1003 (2d Cir. 1995)), *cert. granted*, 142 S. Ct. 1412 (2022).

work.[32] Meanwhile, the creator of the AI system and the dataset assembler responsible for including the copyrighted work in the training dataset might not be directly liable for the volitional conduct of the AI user because they did not prompt the system to create the infringing work.[33]

However, the AI system creator and the dataset assembler may be secondarily liable for the AI user's direct infringement through the doctrines of contributory infringement and/or vicarious liability. Contributory infringement arises when a party "intentionally induc[es] or encourage[es] direct infringement," while vicarious liability occurs when a party "profit[s] from direct infringement while declining to exercise a right to stop or limit" the direct infringement.[34] While actual knowledge of infringing activity would typically support a finding of intent for contributory infringement, courts may find a party liable for contributory infringement even if there is no actual knowledge, but there is willful blindness of the infringing activity.[35] Therefore, given that training datasets for generative AI systems include a significant amount of copyrighted works, the AI creator and the dataset assembler may be liable for contributory infringement of a particular copyrighted work if they have actual knowledge of infringing activity or are willfully blind to such activity.[36] Additionally or alternatively, the AI creator and the dataset assembler may be vicariously liable by profiting from the direct infringement, as the dataset assembler had the ability to control whether a given copyrighted work is within a dataset and the AI creator had the ability to omit the copyrighted work from training but declined to do so. Some AI system creators, such as Stability AI and OpenAI, have implemented copyright infringement notices and takedown procedures to receive the safe-harbor protection from liability provided by section

---

[32] *See* VHT, Inc. v. Zillow Grp., Inc., 918 F.3d 723, 731 (9th Cir. 2019) (discussing the "volitional-conduct requirement" of "direct liability [which] must be premised on conduct that can reasonably be described as the direct cause of the infringement") (quoting Perfect 10, Inc. v. Giganews, Inc., 847 F.3d 657, 666 (9th Cir. 2017)) (emphasis omitted).

[33] *See id.* at 732 (distinguishing between "active and passive participation" for direct infringement) (quoting *Perfect 10, Inc. v. Giganews, Inc.*, 847 F.3d at 667).

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

[35] BMG Rts. Mgmt. (US) LLC v. Cox Commc'ns, Inc., 881 F.3d 293, 310 (4th Cir. 2018) (holding that "at least willful blindness" if not actual knowledge is required to prove contributory infringement); Luvdarts, LLC v. AT & T Mobility, LLC, 710 F.3d 1068, 1073 (9th Cir. 2013) ("Willful blindness of specific facts would establish knowledge for contributory liability.").

[36] *See BMG Rts. Mgmt.* 881 F.3d at 311–12 (4th Cir. 2018). To be clear, liability for contributory infringement arises requires actual knowledge of or willful blindness to *specific* infringing activity, rather than a general awareness that infringing activity is occurring or possible.

512 of the Digital Millennium Copyright Act (DMCA).[37] While these procedures might be helpful for identifying and addressing specific instances of infringing activity, it seems unlikely that the safe harbor of section 512 itself will shield AI system creators from liability because the service they provide (i.e., access to their generative AI systems) is generally inconsistent with the definition of "service provider" in the statute.[38]

Further, the creators of AI systems and dataset assemblers themselves may be found to directly infringe the exclusive rights in copyrighted works. For example, an AI creator may directly infringe during training by reproducing a copyrighted work (e.g., in memory) for input to the system, and possibly by creating intermediate copies during training itself, depending on the nature of training and whether the intermediate copies are fixed.[39] In view of the AI creator's direct infringement, the dataset assembler may be secondarily liable for including the copyrighted work in the dataset.

A dataset assembler may directly infringe by reproducing a copyrighted work in a dataset and distributing the dataset containing the copyrighted work. A dataset assembler directly infringes during assembly by creating copies of copyrighted works and distributing such works via the dataset. Some dataset assemblers, such as LAION, however, seek to sidestep direct infringement by creating datasets that only contain metadata relating to copyrighted works, such as URLs.[40]

Furthermore, if direct infringement by an AI user occurs, the AI system creator and the dataset assembler could potentially be found directly liable for the infringing work, despite not directly prompting the generative AI system themselves, if there is evidence that they actively

---

[37] *See, e.g.*, *Dream Studio Terms of Service*, DREAMSTUDIO, https://beta.dreamstudio.ai/terms-of-service, (Mar. 3, 2023); *Terms of Use,* OPENAI (Nov. 14, 2023), https://openai.com/policies/terms-of-use.

[38] *See* 17 U.S.C. § 512(k)(1) (defining "service provider" as "an entity offering the transmission, routing, or providing of connections for digital online communications, between or among points specified by a user, of material of the user's choosing, without modification to the content of the material as sent or received"). However, to the extent that AI system creators allow AI users to upload content to train the pre-trained generative AI systems, AI system creators might be able to rely on section 512 for safe harbor protection from any infringement liability that would arise through such actions.

[39] *See supra* note 20.

[40] Beaumont, *supra* note 25 (discussing the LAION-5B dataset as "a large-scale dataset for research purposes" and noting that the dataset, which only contains image metadata rather than the images themselves, is licensed under the Creative Commons CC-BY 4.0 license).

"selected" the copyrighted works being infringed.[41] In other words, the direct infringement of a specific copyrighted work by the AI system creator and the dataset assembler might provide a "nexus" to the direct infringement of the same copyrighted work by the AI user.[42]

II. IS THERE A LICENSE FOR THAT?

A license, whether express or implied, to use a copyrighted work is a defense to a claim of copyright infringement.[43] A party accused of copyright infringement might overcome the infringement claim by establishing that they possess a license to the copyrighted work.[44] In response to the accused party showing a license that allegedly excuses the infringement, however, the copyright owner can demonstrate that the accused party's conduct exceeded the scope of the license.[45] Therefore, "[t]o prevail on a claim of copyright infringement, a plaintiff must prove ownership of a copyright and a copying of protectable expression beyond the scope of [a] license."[46]

Datasets are often formed from large repositories of copyrighted material into which the copyright owners uploaded their works for lawful distribution, but the user agreements—both for users uploading their works and for users who may be accessing the material—may or may not address use for dataset inclusion.[47] The widespread use of standard

---

[41] VHT, Inc. v. Zillow Grp., Inc., 918 F.3d 723, 732 (quoting Perfect 10, Inc. v. Giganews*, Inc.*, 847 F.3d 657, 670 (9th Cir. 2017)).

[42] *See id.* (noting that for direct infringement, "[t]here must be actual infringing conduct with a nexus sufficiently close and causal to the illegal copying that one could conclude that the machine owner himself trespassed on the exclusive domain of the copyright owner") (quoting CoStar Grp., Inc. v. LoopNet, Inc., 373 F.3d 544, 550 (4th Cir. 2004)).

[43] Effects Assocs., Inc. v. Cohen, 908 F.2d 555, 559 (9th Cir. 1990) (finding an implied license to special effects footage based on conduct that overcame a copyright infringement claim).

[44] *See, e.g.*, Rimini St., Inc. v. Oracle Int'l Corp., 473 F. Supp. 3d 1158, 1204–05 (D. Nev. 2020).

[45] *Id.* at 1204.

[46] *Id.* at 1202 (quoting MAI Sys. Corp. v. Peak Computer, Inc., 991 F.2d 511, 517 (9th Cir. 1993)).

[47] BookCorpus is one such dataset that has been used to train OpenAI's GPT-N models. Jack Bandy & Nicholas Vincent, *Addressing "Documentation Debt" in Machine Learning Research: A Retrospective Datasheet for BookCorpus*, ARXIV 1 (May 11, 2021, 5:59 PM), https://arxiv.org/abs/2105.05241; Yukun Zhu et al., *Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books*, ARXIV 1, 2 (June 22, 2015, 7:26 PM), https://arxiv.org/abs/1506.06724 (discussing the creation of the BookCorpus dataset as well as the MovieBook dataset which includes a number of famous books along with time-stamped subtitles for movies adapted from those books). BookCorpus was

intellectual property contract provisions in user agreements may result in licenses that are largely unread or respected by different parties, which may push the limits of the interface of federal intellectual property law and state contract law.[48]

In order to evaluate how liability may be assigned in the event of copyright infringement with a generative AI system, the various agreements between the entities potentially involved (i.e., the copyright owners, the Internet platforms, the dataset assemblers, the AI creators, and the end users of the generative AI systems) should be evaluated. Specifically, if licenses to copyrighted works are established through these agreements, then such licenses may provide an affirmative defense to a claim of copyright infringement. However, even if an agreement purports to establish a license, further questions regarding the scope of the license and even the validity or enforceability of certain provisions of the license remain.

A. Untangling the Web of Agreements

1. Agreements Between Copyright Owners and Internet Platforms

The agreements between copyright owners and Internet platforms are, without much exception, the platforms' terms of use.[49] These agreements, sometimes called "clickwrap" or "browsewrap" agreements depending on how much engagement they require from users but not depending on whether users actually read them, are drafted by the

formed from copyrighted books uploaded by their authors to Smashwords for distribution, but none of the Terms of Service for Hugging Face (the dataset repository hosting BookCorpus) or Smashwords, nor the brief nod to licensing on the BookCorpus page indicate that the use of the copyrighted works for training generative AI systems is expressly permitted. *See Datasets: bookcorpus*, HUGGING FACE, https://huggingface.co/datasets/bookcorpus (last accessed Feb. 8, 2023); *Terms of Service*, HUGGING FACE (Sept. 15, 2022), https://huggingface.co/terms-of-service; *Terms of Service*, SMASHWORDS, https://www.smashwords.com/about/tos (May 22, 2023).

[48] *See generally* Amit Elazari Bar On, *Unconscionability 2.0 and the IP Boilerplate: A Revised Doctrine of Unconscionability for the Information Age*, 34 BERKELEY TECH. L.J. 567-68 (2019), https://doi.org/10.15779/Z38PG1HP01 (discussing issues with the use of "IP boilerplate" and proposing an unconscionability framework to address them); Lydia Pallas Loren, *Slaying the Leather-Winged Demons in the Night: Reforming Copyright Owner Contracting with Clickwrap Misuse*, 30 OHIO N.U. L. REV. 495, 499-500 (2004) (discussing abuses by copyright owners through adhesive standard contracts).

[49] Alternatively referred to as terms of service.

platforms and, therefore, tend to favor the platforms.[50] These standard contracts are not negotiated and tend to include standard "boilerplate" terms for intellectual property.[51]

As one example, the terms of service for an Internet platform usually include a provision expressly granting rights to user-uploaded content, such as a license to the platform and a license to other users of the platform to reproduce, distribute, display, perform, and even create derivative works.[52] For an Internet platform whose business model relies on distributing content created by users to other users—which is to say, just about every Internet platform—such an agreement to license the user's exclusive rights in their uploaded content to the platform is a reasonable bargain.[53] Some licenses expressly grant the platforms a right to sublicense the copyrighted work and even specify that the license is "perpetual" and "irrevocable."[54] Regarding the uploaded content, terms of service also typically assign responsibility for the content to the uploading user and require that the uploading user be the copyright

---

[50] Kathleen C. Riley, *Data Scraping as a Cause of Action: Limiting Use of the CFAA and Trespass in Online Copying Cases*, 29 FORDHAM INTELL. PROP. MEDIA & ENT. L.J. 245, 264 (2019) ("A clickwrap license is an agreement that goes into effect when a website user is offered terms and conditions and clicks 'I agree,' while browserwrap licenses are terms and conditions that a user is said to have agreed to by virtue of using an application ('app') or website."); Amit Elazari Bar On, *supra* note 48, at 576 ("Although widespread in both virtual and non-virtual realms, these contracts usually remain hidden on a deserted web page that creators and users never read. In some cases, they take the form of a clickwrap agreement that users spend less than one second reading before they click on so they can use the 'free' service of the platform."); *Id.* at 653 (discussing study of 647 end-user licenses that indicated a bias in favor of the software companies who drafted them).

[51] *See* Amit Elazari Bar On, *supra* note 48, at 589–91. As far as adhesive standard contracts go, terms of service contracts are distinct from EULAs because terms of service are drafted by the party who does not own the intellectual property, while EULAs are drafted by the party who does. *Id.* at 584.

[52] *See, e.g.*, *Terms of Service*, YOUTUBE (Jan. 5, 2022), youtube.com/t/terms?archive=20220105 ("You retain ownership rights in your Content. However, we do require you to grant certain rights to YouTube and other users of the Service … ."); *Reddit User Agreement*, REDDIT, https://www.redditinc.com/policies/user-agreement-september-12-2021 (last updated Aug. 12, 2021) ("When Your Content is created with or submitted to the Services, you grant us a worldwide, royalty-free, perpetual, irrevocable, non-exclusive, transferable, and sublicensable license to use, copy, modify, adapt, prepare derivative works of, distribute, store, perform, and display Your Content . . . ."); *Terms of Service*, SMASHWORDS, https://www.smashwords.com/about/tos (last updated June 9, 2022) ("The Author hereby grants and assigns to Smashwords the non-exclusive worldwide right to digitally publish, distribute, market and sell ("Publish"), and to license others to do so, the work identified on the front page of your submission . . .").

[53] Such a license assumes that the user created the content and therefore owns the copyright in the content.

[54] *See Reddit User Agreement, supra* note 52.

owner.[55] The terms of service for Hugging Face, an Internet platform that hosts user-uploaded datasets and machine learning models, include similar provisions.[56] Thus, the agreements between content creators (i.e., copyright owners) and the content-hosters (i.e., Internet platforms) tend to include broad intellectual property grants to the platforms, including rights to sublicense, for any copyrighted works uploaded to the platform.

In addition to the standard assignment of responsibility and liability for any infringement claims to the users, terms of service often include indemnity provisions and limitations of liability that cap the value of liability. However, an issue with these contractual attempts to limit liability may arise when liability limitations conflict with indemnity provisions.[57] For example, the terms of service for a public dataset repository may include a provision limiting the aggregate liability for each party to the other to some nominal amount, which could severely limit the effectiveness of an indemnity provision.[58] This type of conflict

---

[55] *Terms of Service*, YOUTUBE (Jan. 5, 2022), https://www.youtube.com/static?template=terms ("[T]he Content you submit must not include third-party intellectual property (such as copyrighted material) unless you have permission from that party or are otherwise legally entitled to do so. You are legally responsible for the Content you submit to the Service."); *Reddit User Agreement*, REDDIT, https://www.redditinc.com/policies/user-agreement-september-12-2021 (last updated Aug. 12, 2021) ("By submitting Your Content to the Services, you represent and warrant that you have all rights, power, and authority necessary to grant the rights to Your Content contained within these Terms. Because you alone are responsible for Your Content, you may expose yourself to liability if you post or share Content without all necessary rights."); *Terms of Service*, SMASHWORDS, https://www.smashwords.com/about/tos (last updated June 9, 2022) ("If you upload (publish) a work to Smashwords, you understand and warrant that you are the legal publisher of this work; you control all rights and assume all liabilities associated with the publication of this work; and you warrant and affirm that no aspect of your work infringes or violates the rights of another person, party or entity.").

[56] *See Terms of Service*, HUGGING FACE (Sept. 15, 2022), https://huggingface.co/terms-of-service ("You are **solely responsible for the Content you post**, publish, display or otherwise make available on our Website . . . . You represent and warrant that you have ownership, control, and responsibility for the Content you post or otherwise make available on our Website, or otherwise have the right to do so. Your Content must not . . . infringe or misappropriate any rights of any person or entity.").

[57] *See* Geoff Sutcliffe, *When the Limitation of Liability Is Not So Limiting*, A.B.A.: LANDSLIDE EXTRA (June 30, 2021), https://www.americanbar.org/groups/intellectual_property_law/publications/landslide-extra/limitations-of-liability/.

[58] *Terms of Service*, HUGGING FACE (Sept. 15, 2022), https://huggingface.co/terms-of-service ("Either Party's . . . aggregate liability to the other Party or any third party in any circumstance will not exceed the amount that you paid us during the 12-month period immediately preceding the last claim . . . you agree to **indemnify**, defend and hold harmless **us and Related Parties** from all claims, liability, and expenses, including attorney's fees, arising out or in connection with your **use** of . . . the Services . . .").

can easily occur when platforms and users take boilerplate provisions for granted.

Thus, there is typically a license agreement between the copyright owners and the Internet platforms.[59] Internet platforms are free from assertions of infringement liability from their users because the scope of these licenses is broad. Despite the breadth of the licenses, they are generally non-exclusive, so Internet platforms do not have standing to assert the exclusive rights of copyright owners.[60]

2. Agreements Between Internet Platforms and Dataset Assemblers

Absent a particular negotiated agreement between a dataset assembler and an Internet platform, the operating agreement between these parties also tends to be the platform's terms of service.[61] The terms of service will sometimes grant a non-exclusive license to users of an Internet platform to access user-provided content hosted on the platform.[62] Therefore, this license authorizes the platform to exercise the copyright owner's exclusive rights, such as reproducing, distributing, displaying, and performing copyrighted works, enabling the platform to facilitate user access.

Further, the terms of use for Internet platforms that host user-uploaded content typically include provisions prohibiting certain behaviors by users accessing the platform, such as prohibitions on

---

[59] If a user uploads copyrighted content that the user does not own, an Internet platform complying with the take-down procedures of section 512 may be able to receive safe harbor protection from infringement liability, unless the Internet platform is using the copyrighted content to train a generative AI system. *See supra* text accompanying notes 37–38.

[60] HyperQuest, Inc. v. N'Site Sols., Inc., 632 F.3d 377, 382 (7th Cir. 2011) ("The corollary to [the rule that assignee of an exclusive right has standing] is that a person holding a non-exclusive license is not entitled to complain about any alleged infringement of the copyright."); Riley, *supra* note 50, at 308 ("The lack of standing in copyright infringement lawsuits for user-based services explains why services like Facebook and LinkedIn have resorted to the CFAA as a potential remedy for copying of their websites.").

[61] *See supra* note 52.

[62] *See, e.g.*, *Terms of Service*, YOUTUBE (Jan. 5, 2022), youtube.com/t/terms?archive=20220105 ("You also grant each other user of the Service a worldwide, non-exclusive, royalty-free license to access your Content through the Service, and to use that Content, including to reproduce, distribute, prepare derivative works, display, and perform it, only as enabled by a feature of the Service (such as video playback or embeds). For clarity, this license does not grant any rights or permissions for a user to make use of your Content independent of the Service.").

reproducing or distributing content available on the platform.[63] These prohibitions on reproduction or distribution essentially echo the exclusive rights in reproduction and distribution afforded to owners of copyrighted works.[64] As discussed above, standard terms of service for Internet platforms only provide a non-exclusive license from the copyright owner, but "the holder of a nonexclusive license may not sue others for infringement."[65] Thus, an Internet platform seeking to enforce these terms against a dataset assembler who may be violating the terms by reproducing and distributing the content in a dataset could not bring a copyright infringement suit and instead would have to rely on a breach of contract. However, such state contract law claims may be preempted by section 301 of the Copyright Act, so even a breach-of-contract claim may be unavailable.[66]

In addition to boilerplate prohibitions that duplicate exclusive rights in copyrighted works, terms of service often include prohibitions on automated access to the platform.[67] These prohibitions on automated access are intended to prevent the scraping of content hosted on the platform.[68] On its face, these prohibitions would suggest that the automated scraping of content across the Internet by dataset assemblers

---

[63] *See, e.g.*, *Terms of Service*, YOUTUBE (Jan. 5, 2022), youtube.com/t/terms?archive=20220105 ("You are not allowed to . . . access, reproduce, download, distribute, transmit, broadcast, display, sell, license, alter, modify or otherwise use any part of the Service or any Content except: (a) as expressly authorized by the Service; or (b) with prior written permission from YouTube and, if applicable, the respective rights holders . . . ."); *Reddit User Agreement*, REDDIT, https://www.redditinc.com/policies/user-agreement-september-12-2021 (last updated Aug. 12, 2021) ("Except and solely to the extent such a restriction is impermissible under applicable law, you may not, without our written agreement: license, sell, transfer, assign, distribute, host, or otherwise commercially exploit the Services or Content; modify, prepare derivative works of, disassemble, decompile, or reverse engineer any part of the Services or Content . . ."); *Terms of Service*, SMASHWORDS, https://www.smashwords.com/about/tos (last updated June 9, 2022) ("While using the Site, Services or Work, End Users agree to not: . . . copy, modify, or distribute content from the Site . . .").

[64] *See* 17 U.S.C. §106(1), (3).

[65] Riley, *supra* note 50, at 307 (quoting Davis v. Blige, 505 F.3d 90, 101 (2d Cir. 2007)).

[66] *See infra* Section III(A)(4) (discussing circuit split regarding whether the Copyright Act preempts breach-of-contract claims). This further explains why Internet platforms typically resort to the Computer Fraud and Abuse Act (CFAA) as a potential remedy.

[67] *See, e.g.*, *Terms of Service*, YOUTUBE (Jan. 5, 2022), youtube.com/t/terms?archive=20220105 ("You are not allowed to . . . access the Service using any automated means (such as robots, botnets or scrapers) except (a) in the case of public search engines, in accordance with YouTube's robots.txt file; or (b) with YouTube's prior written permission . . . ."); Riley, *supra* note 50, at 257 (discussing the inclusion of anti-automated access provisions in the terms of service of a large variety of Internet platforms).

[68] *See* Riley, *supra* note 50, at 257–59.

would be unauthorized and unlawful.

Courts have hesitated to find such scraping unlawful when these no-automated-access provisions are litigated. For example, in *HiQ Labs, Inc. v. LinkedIn Corp.*, the Ninth Circuit on remand from the Supreme Court affirmed a preliminary injunction "forbidding LinkedIn from denying hiQ access to publicly available LinkedIn member profiles."[69] The data analytics company hiQ used automated bots to scrape data from publicly available LinkedIn member profiles, which hiQ then used in its analytics products.[70] Given LinkedIn's efforts to prevent automated scraping through technological tools as well as its User Agreement, LinkedIn sent a cease-and-desist letter to hiQ asserting potential violations of the Computer Fraud and Abuse Act (CFAA) and the Digital Millennium Copyright Act (DMCA).[71] While some interpret the court's affirmation of the preliminary injunction as allowing data scraping of publicly available information,[72] the hiQ court relied on a narrow interpretation of the CFAA to reach its decision and noted that "other causes of action, such as copyright infringement" may still be applicable.[73]

Thus, while there is an express agreement between Internet platforms and users, which includes dataset assemblers, some actions such as web scraping of content from a platform may exceed the scope of the agreement.[74] In the event that web scraping exceeds the scope of a license between the Internet platform and a dataset assembler, such that the dataset assembler's conduct is unauthorized, the dataset assembler may hope to argue that the conduct is fair use.[75] However, fair use is only

---

[69] HiQ Labs, Inc. v. LinkedIn Corp., 31 F.4th 1180, 1184–85 (9th Cir. 2022).
[70] *Id.* at 1187.
[71] *Id.* at 1186–87.
[72] *See* Jennifer Oliver, *Ninth Circuit Holds Data Scraping is Legal in hiQ v. Linkedin*, CAL. LAWS. ASS'N (May 2022), https://calawyers.org/privacy-law/ninth-circuit-holds-data-scraping-is-legal-in-hiq-v-linkedin; Camille Fischer & Andrew Crocker, *Victory! Ruling in* hiQ v. LinkedIn *Protects Scraping of Public Data*, ELEC. FRONTIER FOUND. (Sept. 10, 2019), https://www.eff.org/deeplinks/2019/09/victory-ruling-hiq-v-linkedin-protects-scraping-public-data.
[73] HiQ Labs, Inc. v. LinkedIn Corp., 31 F.4th at 1180, 1201 (9th Cir. 2021). However, the court hinted at the impropriety of LinkedIn controlling access to "data that the companies [hiQ and LinkedIn] do not own." *Id.* at 1202.
[74] *See, e.g.*, Sun Microsystems, Inc. v. Microsoft Corp., 81 F. Supp. 2d 1026, 1032 (N.D. Cal. 2000) (distinguishing contractual covenants from "conditions of, or restrictions on, the license grants" in a software license).
[75] *See, e.g.*, Perfect 10, Inc. v. Amazon.com, Inc., 508 F.3d 1146, 1155, 1166 (9th Cir. 2007) (finding that Google's database of webpages and database of image thumbnails built from automatically accessing or crawling the Internet was fair use of copyrighted works because of the "significantly transformative nature of Google's search engine"); Kelly v. Arriba Soft Corp., 336 F.3d 811, 816, 818 (9th Cir. 2003) (finding that a search

an affirmative defense to a copyright infringement claim, which the Internet platform cannot bring if they only have a non-exclusive license to their hosted content.[76]

There is the possibility that an Internet platform and a dataset assembler could negotiate a license agreement for the content. This agreement would be prudent if the platform has an expressly granted right to sublicense the content to other parties because the Internet platform is therefore free to expressly sublicense the copyrighted works in bulk to a dataset assembler.[77] Even in the absence of an expressly granted sublicense to a dataset assembler, if express non-exclusive licenses from copyright owners to an Internet platform are valid and enforceable, the conduct between the Internet platform and the dataset assembler could imply intent to sublicense and thus give rise to an implied sublicense for the copyrighted works to the dataset assembler.[78]

3. Agreements Between Dataset Assemblers and AI Creators

Various types of agreements may exist between dataset assemblers and AI creators. These agreements may range from private, negotiated licenses with specifically tailored terms, to public, standard adhesive contracts such as EULAs or terms of service.[79] One option

---

engine's database of copied images scraped without authorization from other websites authorized to display the images was fair use because the display of the copied images as low-resolution thumbnails was a "transformative" purpose). *But see id.* at 819 ("Arriba's use of the images serves a different function than Kelly's use—improving access to information on the internet versus artistic expression."); Associated Press v. Meltwater U.S. Holdings, Inc., 931 F.Supp.2d 537, 544, 561 (S.D.N.Y. 2013) (holding that a database of scraped news articles from webpages to provide news excerpts was not fair use because the use was not transformative and "[e]xploitation of search engine technology to gather content does not answer the question of whether the business itself functions as a search engine.").

[76] *See supra* text accompanying note 65.

[77] *See infra* Section III(A)(2).

[78] *See* Photographic Illustrators Corp. v. Orgill, Inc., 953 F.3d 56, 58, 62–64 (adopting a flexible approach to determine whether a non-exclusive licensee "sufficiently manifested an intent to grant . . . a sublicense"); RESTATEMENT OF COPYRIGHT §27 cmt. h (AM. L. INST. Tentative Draft No. 3, 2022) (discussing implied authorization for non-exclusive licensees to sublicense); *infra* text accompanying notes 153-156 (discussing circuit split on implied rights to sublicense).

[79] These licenses are premised on the copyrightability of the dataset itself, which would require at least a "modicum of creativity" to satisfy the originality requirement for a copyright. Feist Publ'ns, Inc. v. Rural Tel. Serv. Co., 499 U.S. 340, 362 (1991) (holding that the "selection, coordination, and arrangement" of listings in a phone book were not sufficiently original to warrant copyright protection). This paper sets aside the question of the copyrightability of datasets because most datasets used for training generative AI systems have had at least some thoughtful effort involved in the

within this spectrum are the Creative Commons licenses, which allow copyright owners to specify the level of protection they seek for their works.[80] For example, the LAION-5B image dataset is licensed to AI creators under the Creative Commons Attribution 4.0 license (CC-BY 4.0).[81] Creative Commons licenses, including CC-BY 4.0, explicitly note that the license agreement does not apply to uses that qualify as fair use.[82] Regarding scope, a Creative Commons license provides a grant of a "worldwide, royalty-free, non-sublicensable, non-exclusive, irrevocable license to exercise the Licensed Rights in the Licensed Material to . . . reproduce and Share the Licensed Material, in whole or in part; and . . . produce, reproduce, and Share Adapted Material."[83] Further, a Creative Commons license "offers the Licensed Material as-is and as-available, . . . makes no representations or warranties of any kind concerning the Licensed Material," and limits all liability "[t]o the extent possible . . . on any legal theory."[84]

As another example of a dataset license, the Common Crawl dataset is one of the datasets used to train OpenAI's GPT models.[85] This dataset includes petabytes of data collected during monthly comprehensive web crawls.[86] By systematically scraping the entire Internet, the Common Crawl dataset naturally includes a significant amount of copyrighted material.[87] The Common Crawl Foundation only grants a "non-assignable, non-transferable, non-sublicensable limited

---

selection, coordination, and arrangement in order to improve training performance. *See* Brown et al., *supra* note 17, at 8–9. That said, originality may be a valid inquiry for datasets assembled from indiscriminate web scraping, such as the Common Crawl dataset.

[80] *See About CC Licenses*, CREATIVE COMMONS, https://creativecommons.org/share-your-work/cclicenses/ (last visited Mar. 21, 2023).

[81] Romain Beaumont, *LAION-5B: A New Era of Open Large-Scale Multi-Modal Datasets*, LAION (Mar. 31, 2022), https://laion.ai/blog/laion-5b/ (discussing the LAION-5B dataset as "a large-scale dataset for research purposes" and noting that the dataset, which only contains image metadata rather than the images themselves, is licensed under the Creative Commons CC-BY 4.0 license).

[82] *Creative Commons by 4.0 Deed: Attribution 4.0 International*, CREATIVE COMMONS, https://creativecommons.org/licenses/by/4.0/ (last visited Mar. 21, 2023) ("You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable exception or limitation. The rights of users under exceptions and limitations, such as fair use and fair dealing, are not affected by the CC licenses.").

[83] *Creative Commons by 4.0 Legal Code: Attribution 4.0 Int'l*, CREATIVE COMMONS, § 2(a)(1) https://creativecommons.org/licenses/by/4.0/legalcode .

[84] *Id.* § 5.

[85] *See* Brown et al., *supra* note 17, at 8–9.

[86] *Overview*, COMMON CRAWL, https://commoncrawl.org/overview (last visited May 28, 2024).

[87] *Id.* (It contains raw web page data, extracted metadata, and text extractions.).

license to access the [s]ervice . . . ."[88] The limited license includes use restrictions such as, among other things, "violating the rights of another individual or entity, including but not limited to such party's intellectual property rights or other proprietary rights."[89] The terms of use explicitly disclaims any warranties regarding the scraped content, and limits liability for Common Crawl.[90]

As another example of a training dataset, vAIsual provides datasets that are not only "ethically clean" datasets "respect[ing] copyright of creators and personal rights of models", but also "legally clean" because all of the rights to the data in the datasets are licensed and cleared.[91] Despite these public assurances, the actual license agreements for vAIsual's datasets nevertheless disclaim any warranties and provide no guarantees.[92] Regarding scope, the license granted is "a personal, non-exclusive, non-sublicensable and non-transferable, limited, revocable license to use the Dataset."[93] The license includes a notable restriction on "the creation of Synthetic Media . . . that is generated through artificial production, manipulation, or modification by automated means, including but not limited through the use of artificial intelligence algorithms."[94] This use restriction is likely targeting the use of the dataset for creating "deepfake" videos where the likeness of one person is replaced by another person,[95] but the license's definition of Synthetic Media is broad enough to cover output of generative AI systems. In addition to this use restriction, the agreement also specifies that "[t]he Dataset shall be used exclusively for machine learning purposes," and that "[d]atasets can be used by the End User for any machine learning purposes and training neural networks for any kind of any application."[96] The agreement for this "legally clean" dataset both refuses to legally

---

[88] *Common Crawl Foundation – Terms of Use*, COMMON CRAWL, https://commoncrawl.org/terms-of-use (last visited Aug. 21, 2023).
[89] *Id.* § 2(d).
[90] *Id.* §§ 3, 6, 7.
[91] DATASET SHOP, https://www.datasetshop.com (last visited Mar. 20, 2023).
[92] *Standard License (One-Time Purchase)*, DATASET SHOP, https://www.datasetshop.com/standard-license-one-time-purchase (last visited Mar. 20, 2023) ("The Dataset made available is provided 'as is' without vAIsual's warranty of any kind, either express or implied, including, but not limited to, any implied warranty against infringement of third parties' rights including but not limited to Intellectual Property Rights").
[93] *Id.* § 1.1.
[94] *Id.*
[95] *See* David Gray Widder, Dawn Nafus, Laura Dabbish & James Herbsleb, *Limits and Possibilities for "Ethical AI" in Open Source: A Study of Deepfakes*, 2022 ACM CONF. ON FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY, June 21–24, 2022, SEOUL, REP. OF KOR. (2022), at 2–3.
[96] *Standard License*, *supra* note 93, §§ 2.1–2.2.

represent and warrant the cleanliness of the dataset, and simultaneously restricts and allows the use of the dataset for training generative AI systems. Given the uncertainty about whether the license permits using the dataset to train generative AI systems, a licensee may have to rely on fair use if the usage is found to be outside the scope of the license. The vAIsual license agreement also limits its liability for the use of "the Dataset or the results received from use of the Dataset."[97]

In general, these various licenses between dataset assemblers and AI creators tend to provide limited scope to use the datasets. This limited scope is likely due to the fact that these licenses are drafted by the parties who ostensibly own the copyrighted work being licensed—the dataset itself[98]—so the licenses are biased towards the drafting party. Insofar that the use of the dataset is contemplated, the licenses range from providing broad reproduction and distribution rights with express acknowledgement of fair use for any unauthorized use,[99] to providing conflicting restrictions and allowances on the same use. This suggests a reliance on fair use doctrine to address the more complicated issues surrounding generative AI.[100]

4. Agreements Between AI Creators and AI Users

Similarly, the user agreements between end users of generative AI systems and the AI creators attempt to limit liability by forbidding particular behavior that would result in litigation. For example, the terms of use for products released by OpenAI, including the generative AI systems ChatGPT and Dall-E, grant users "a non-exclusive right to use[] the Services in accordance with these Terms," and are clear regarding copyright infringement: "[y]ou may not . . . use [the] Services in a way that infringes, misappropriates or violates anyone's rights."[101] The OpenAI terms further include boilerplate provisions for indemnification and limitation of liability, as well as a complete "as-is" disclaimer of any

---

[97] *Id.* § 8.1.

[98] The dataset as a compilation of data is one type of copyrighted work, but all of the copyrighted works contained in that compilation are owned by someone else. *See* 17 U.S.C. § 101 ("A 'compilation' is a work formed by the collection and assembling of preexisting materials or of data that are selected, coordinated, or arranged in such a way that the resulting work as a whole constitutes an original work of authorship.").

[99] *See Creative Commons Attribution 4.09 International Public License*, *supra* note 83.

[100] *See supra* notes 96-97 and accompanying text.

[101] OPENAI, *Terms of Use* (Nov. 14, 2023), https://openai.com/policies/terms-of-use. Ironically, there is also a prohibition on using any automated or programmatic method to extract data or output from the Services, including scraping, web harvesting, or web data extraction. *Id.*

warranties.[102]

The license from Stability AI for the use of the generative AI system Stable Diffusion is an Open Responsible AI License (Open RAIL).[103] This license provides a generous grant consisting of a "perpetual, worldwide, non-exclusive, no-charge, royalty-free, irrevocable copyright license to reproduce, prepare, publicly display, publicly perform, sublicense, and distribute" the model, derivatives of the model, and software materials used to define and train the model.[104] The license further specifies that "[l]icensor claims no rights in the Output You generate using the Model," and that "[y]ou are accountable for the Output you generate and its subsequent uses."[105] The license also requires users "not to use the Model . . . [i]n any way that violates any applicable national, federal, state, local, or international law or regulation" which includes, as just one example, violations of the federal copyright laws.[106] Much like other licenses, the Open RAIL also disclaims any warranty while providing the AI system on an "as-is" basis, and limits the liability of the licensor.[107]

The license provided by Stability AI is one of a variety of "Responsible AI" licenses developed by the RAIL Initiative for the purpose of encouraging open sharing of AI technology, while restricting such technology from being used in "harmful applications."[108] These RAILs are specifically constructed to impose restrictions on certain "behavioral uses" and are inspired by the ethos of open-source license models.[109] By including specific use restrictions, licensors such as AI system creators may turn to the courts to enforce the license terms against licensees (i.e., AI users) upon breach. Enforcing the license

---

[102] *Id.* § 7.
[103] Rombach et al., *High-Resolution Image Synthesis with Latent Diffusion Models in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2022), https://huggingface.co/stabilityai/stable-diffusion-2-1; *CreativeML Open RAIL++-M License*, HUGGING FACE (Nov. 24, 2022), https://huggingface.co/stabilityai/stable-diffusion-2/blob/main/LICENSE-MODEL.
[104] *CreativeML Open RAIL++-M License*, *supra* note 103, § II(2). The license also grants a similarly generous patent license. *Id.* § II(3).
[105] *Id.* § III(6).
[106] *Id.* attach. A.
[107] *Id.* §§ IV(9)–(10).
[108] *About*, RESPONSIBLE AI LICENSES, https://www.licenses.ai/about (last visited Mar. 22, 2023); Danish Contractor et al., *From RAIL to Open RAIL: Topologies of Rail Licenses*, RESPONSIBLE AI LICENSES (Aug. 18, 2022), https://www.licenses.ai/blog/2022/8/18/naming-convention-of-responsible-ai-licenses (describing the various RAILs in comparison to Open Source and Creative Commons terms).
[109] *See* Danish Contractor et al., *Behavioral Use Licensing for Responsible AI*, 2022 ACM CONF. ON FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY, JUNE 21–24, 2022, SEOUL, REP. OF KOR. 778 (2022).

against breaching licensees would function as a deterrence mechanism against AI users who would prefer to "self-enforce" to avoid the risk of expensive litigation."[110] In order to actually deter AI users from using generative AI systems for the restricted uses, which range from violating any laws to providing medical advice,[111] AI system creators should actually enforce these provisions.

B. Locating Infringement Liability in View of License Agreements

Unsurprisingly, every license agreement disclaims liability for the drafting party.[112] When every license agreement in the web of parties disclaims liability, who should be liable for copyright infringement?

1. Structural Features of the Entities May Impact the Analysis

The interplay of various licenses might make identifying liable tortfeasors for copyright infringement extremely fact-specific, where the specific scope of each license must be evaluated to determine whether relevant behavior is authorized by the license.

In some instances, the problem is simplified because one or more of the parties is identical. For example, in *Doe. v. GitHub, Inc.*, a single party—the software code repository GitHub—fulfills the role of Internet platform, dataset assembler, and AI creator.[113] In this case, whether or not GitHub was authorized to use the copyrighted content uploaded by users to train a generative AI system to output software code based on prompts, ultimately comes down to whether such use was within the scope of the license established in the GitHub terms of service.[114]

---

[110] *Id.* at 782. This deterrence mechanism is not taken seriously in the industry, however, where the open release of generative AI models without strong technical safeguards is viewed by some as "problematic." Kyle Wiggers, *This Startup is Setting a DALL-E 2-Like AI Free, Consequences Be Damned*, TECHCRUNCH (Aug. 12, 2022, 1:55 PM), https://techcrunch.com/2022/08/12/a-startup-wants-to-democratize-the-tech-behind-dall-e-2-consequences-be-damned/ ("Doubtless, some of these images are against Stability AI's own terms, but the company is currently relying on the community to flag violations.").

[111] *See CreativeML Open RAIL++-M License*, *supra* note 104, attach. A.

[112] *See supra* §II(A) (discussing license agreements between parties).

[113] *See* Defendants GitHub and Microsoft's Notice of Motions and Motions to Dismiss Operative Complaint in Consolidated Actions at 3, No. 4:22-cv-06823-JST (N.D. Cal. Jan. 26, 2023) ("The version of Codex that powers Copilot was trained on billions of lines of code that GitHub users stored in public GitHub repositories.").

[114] *See id.* at 16–17 ("GitHub's TOS expressly authorizes the training of Copilot.").

In other instances, the problem is complex because the entities create significant space between them and the legal issues. For example, particularly savvy dataset assemblers, such as LAION, have avoided the problem of crawling or web scraping altogether by instead analyzing the Common Crawl dataset.[115] LAION has taken a further step of not including any actual images in their datasets, but rather URLs to the original images.[116] Further, although LAION's efforts to create the LAION-5B dataset were funded and supported by Stability AI, LAION's status as a non-profit German entity provides jurisdictional insulation for LAION against copyright infringement claims.[117] As a result, the copyright infringement claims filed to date are focused on the actions of Stability AI rather than other entities.[118] Nevertheless, relevant license agreements for all parties should be considered when evaluating whether the actions of a given party were authorized.

2. Limited Liability is Not So Limiting

As a general rule, one cannot contract out of liability for tortious behavior.[119] In this light, the widespread limitations on liability provided as a boilerplate term throughout all agreements may not appear to be a strong barrier to assigning liability to a party who has ostensibly disclaimed any liability. An exception to the rule is if there is "a fairly bargained for agreement to limit liability to a reasonable agreed value in

---

[115] *See, e.g.*, *FAQ*, LAION, https://laion.ai/faq/ (last visited Mar. 20, 2023) ("[W]e are not crawling websites to create the datasets. Common Crawl did the crawling part in the past, and they did respect the robots.txt instruction. We only analyse their data and then look at the pictures to assess their value concerning the provided alt text.").
[116] *Id.* ("LAION datasets are simply indexes to the internet, i.e. lists of URLs to the original images together with the ALT texts found linked to those images. While we downloaded and calculated CLIP embeddings of the pictures to compute similarity scores between pictures and texts, we subsequently discarded all the photos. Any researcher using the datasets must reconstruct the images data by downloading the subset they are interested in.").
[117] *See* Wiggers, *supra* note 111 ("[Stability AI CEO and founder Emad] Mostaque says that Stability AI funded the creation of LAION 5B, an open source, 250-terabyte dataset containing 5.6 billion images scraped from the internet."); Complaint at 13, Getty Images (US), Inc. v. Stability AI, Inc., No. 1:23-cv-00135-UNA (D. Del. Feb. 3, 2023) ("Stable Diffusion was trained on 5 billion image-text pairs from datasets prepared by non-party LAION, a Germany entity that works in conjunction with and is sponsored by Stability AI. . . . Stability AI provided LAION with both funding and significant computing resources to produce its datasets in furtherance of Stability AI's infringing scheme."). Claims could potentially be brought under German copyright law, a topic outside the scope of this paper.
[118] *See, e.g.*, Getty Images (US), Inc. v. Stability AI, Inc. Complaint, *supra* note 117.
[119] RESTATEMENT (SECOND) OF CONTRACTS § 195(1) (A.L.I. 1981) ("A term exempting a party from tort liability for harm caused intentionally or recklessly is unenforceable on grounds of public policy."); *id.* §195(2) (similarly for negligently caused harm in certain circumstances).

return for a lower rate."[120] In view of this exception and notwithstanding the blunt disclaimer of any liability, limitations of liability in terms of service tend to specify a liability cap.[121]

The enforceability of these liability limitations may come down to whether or not they are unconscionable, which would occur if the provision is procedurally and substantively unconscionable.[122] Whether the provision is enforceable could be a close call given that most of the agreements are not negotiated and are typically biased towards the drafting party limiting liability.[123]

However, regardless of whether the limitation of liability provisions is enforceable, they are only applicable to the parties to the agreement. Therefore, in situations where the copyright owner is not a party to the infringing party's terms of use, the limitation of liability does not shield the infringing party from liability to the copyright owner for infringement of the copyright owner's exclusive rights in a copyrighted work. Instead, the provision would at most shield the party from a crossclaim, which is when an AI user accused of direct infringement attempts to pursue a claim against the AI creator. Furthermore, the federal copyright laws may preempt attempts to contract out of liability for copyright infringement.[124]

Licenses and user agreements cannot absolve dataset assemblers and AI system creators of their contributory liability for copyright infringement committed by AI users. However, the existence of "substantial non-infringing uses" of an expressive AI system may absolve an AI creator from contributory liability for direct infringement by an AI user.[125] For example, where instances of direct infringement in generated output are relatively anomalous and the generative AI systems are more than capable of substantial non-infringing uses, contributory liability for AI creators may be significantly limited.

III. BEYOND THE FINE PRINT

The potential liability for dataset assemblers and AI creators

---

[120] *Id.* § 195 cmt. a.

[121] *See, e.g.*, OPENAI, *supra* note 102, § 7 ("Our aggregate liability under these terms will not exceed the greater of the amount you paid for the service that gave rise to the claim during the 12 months before the liability arose or one hundred dollars ($100).") (all caps removed).

[122] Meta Platforms, Inc. v. BrandTotal Ltd., 605 F. Supp. 3d 1218, 1253 (N.D. Cal. 2022) (declining to find a limitation of liability unconscionable).

[123] *See supra* notes 48, 50; *see also* Feldman v. Google, Inc., 513 F. Supp. 2d 229, 242 (E.D. Pa. 2007) (holding that a standard clickwrap agreement was not unconscionable because the party had adequate notice and assented to the terms).

[124] *See infra* Section III(A)(4).

[125] Sony Corp. of Am. v. Universal City Studios, Inc., 464 U.S. 417, 442 (1984) (finding no contributory infringement "if the product is widely used for legitimate, unobjectionable purposes. Indeed, it need merely be capable of substantial noninfringing uses").

depends on factors beyond what is in the fine print of the various agreements, such as how courts may interpret the fine print itself and the overall conduct of the parties.

A. Legal Considerations for Drafting License Agreements

As the technology, its impact on society, and copyright law doctrines continue to evolve, a more proactive approach to mitigating risk and respecting the rights of copyright owners may be prudent.

1. Volition or Strict Liability?

Courts often state that "copyright infringement is a strict liability offense" without an intent requirement.[126] Even with a license from a database assembler, an AI creator will still be strictly liable if the dataset includes unauthorized works.[127] However, courts have found that "some element of volition or causation" is necessary to hold a party directly liable for copyright infringement when automated systems are used for the allegedly infringing conduct.[128]

The conduct of an accused infringer in situations involving automated systems, which may occur in the context of dataset assembly and generative AI systems, may be an important consideration. For example, in *VHT, Inc. v. Zillow Group, Inc.*, the Ninth Circuit found Zillow not directly liable for copyright infringement of photos where "[t]he content of the Listing Platform [was] populated with data submitted by third-party sources that attested to the permissible use of that data, and Zillow's system for managing photos on the Listing

---

[126] Brammer v. Violent Hues Prods., 922 F.3d 255, 265 (4th Cir. 2019); *see also* Jacobs v. Memphis Convention & Visitors Bureau, 710 F. Supp. 2d 663, 678 n.21 (W.D. Tenn. 2010) ("Copyright infringement, however, is at its core a strict liability cause of action, and copyright law imposes liability even in the absence of an intent to infringe the rights of the copyright holder."); King Recs., Inc. v. Bennett, 438 F. Supp. 2d 812, 852 (M.D. Tenn. 2006) (quoting Bridgeport Music Inc. v. 11C Music, 154 F. Supp. 2d 1330, 1335 (M.D. Tenn. 2001)) ("Liability for copyright infringement does not turn on the infringer's mental state because 'a general claim for copyright infringement is fundamentally one founded on strict liability.'").

[127] Lemley & Casey, *supra* note 6, at 758 (citing Lipton v. Nature Co., 71 F.3d 464, 471 (2d Cir. 1995)) ("Copyright is a strict liability offense. Acting reasonably in getting a license from the database owner won't help you if the database owner doesn't have a license for each and every one of the hundreds of millions of works, even if they claim they do.").

[128] Religious Tech. Ctr. v. Netcom On-Line Commc'n Servs., Inc., 907 F. Supp. 1361, 1370 (N.D. Cal. 1995) ("Although copyright is a strict liability statute, there should still be some element of volition or causation which is lacking where a defendant's system is merely used to create a copy by a third party.").

Platform was constructed in a copyright-protective way."[129] This finding of no liability is highly instructive. For example, some of the "copyright-protective" behavior used by Zillow to avoid liability included "requir[ing content] providers to certify the extent of their rights to use each photo" and "programm[ing] its automated systems to treat each photo consistently with that scope of use certified to by the third party."[130] Zillow also adopted "trumping" rules to handle content duplicates that favored content with more appropriate rights, and generally "designed its system to avoid and eliminate copyright infringement.[131] For generative AI systems, implementing technical mechanisms that would similarly help avoid copyright infringement, even if imperfectly, could help avoid the volition element required for direct liability.

However, Zillow was held directly liable for copyright infringement with regard to a set of photographs "that were selected and tagged by Zillow moderators for searchable functionality and displayed on [another platform]."[132] Further, the court found that "the searchability function did not constitute fair use."[133] As arguments for fair use sometimes analogize to search engine functionality,[134] this finding of no fair use indicates that dataset assemblers and AI creators should be cautious when refining a massive dataset for training a generative AI system, where a significant amount of selection and even tagging may occur.[135]

2. Scope of Agreements

A breach of a license agreement only results in copyright infringement if the breach exceeds the scope of the license, and if the breach violates an exclusive right.[136] Even if the license is not terminated, the unauthorized behavior outside the scope of the license is actionable if it infringes an exclusive right in a way that is not

---

[129] VHT, Inc. v. Zillow, Inc., 918 F.3d 723, 733 (9th Cir. 2019).
[130] *Id.*
[131] *Id.*
[132] *Id.* at 734.
[133] *Id.*
[134] *See, e.g.*, Authors Guild, Inc. v. Google, Inc., 721 F.3d 132 (2d Cir. 2013); Authors Guild, Inc. v. HathiTrust, 755 F.3d 87 (2d Cir. 2014); Perfect 10, Inc. v. Amazon.com, Inc., 508 F.3d 1146 (9th Cir. 2007).
[135] *See* Brown et al., *supra* note 17, at 8–9.
[136] MDY Indus. v. Blizzard Ent., Inc., 629 F.3d 928, 939–41 (9th Cir. 2010), *as amended on denial of reh'g* (Feb. 17, 2011), *opinion amended and superseded on denial of reh'g*, No. 09-15932, 2011 WL 538748 (9th Cir. Feb. 17, 2011) (distinguishing covenants from conditions in a license agreement).

authorized by the license.

The use of data for the purposes of training and implementing generative AI systems may qualify as a new use and exceed the scope of an existing license. In *Random House, Inc. v. Rosetta Books LLC*, the district court held that exclusive license agreements between authors and Random House for publishing books did not cover the publication of ebooks. Consequently, since Random House did not own the exclusive right to publish ebooks, they were unable to prove that Rosetta Books's publication of ebooks infringed their copyrights.[137] In making their decision, the court relied on "the language of the license contract and basic principles of interpretation" to conclude that the license grant to "print, publish and sell the work in book form" was limited to known analog formats and did not extend to new digital formats that enabled different ways to interact with the work, such as through manipulating the text.[138] Similarly, the language of existing licenses to copyrighted works contained in Internet platforms' terms of service may be limited to facilitating user access to the copyrighted works via the Internet platforms.[139] These licenses may not extend to using the copyrighted works for training generative AI systems if such use is "so dissimilar" or "not analogous" to the uses contemplated by the licenses as to qualify as a "new use."[140]

The question whether the scope of a license will cover training a generative AI system despite the license not expressly discussing such use may soon be addressed by the courts. In ongoing class-action litigation, *Doe v. Github, Inc.*, two generative AI systems, Codex and Copilot, trained to generate software code are accused of "output[ting] copyrighted materials without following the terms of the applicable licenses."[141] Codex, developed by OpenAI, and Copilot, jointly developed by OpenAI and GitHub, were "each trained on a large corpus of publicly accessible software code and other materials."[142] Plaintiffs accuse these systems of outputting "a near-identical reproduction of code from the training data."[143] Some of the publicly accessible software code was allegedly licensed under certain open source and Creative Commons

---

[137] Random House, Inc. v. Rosetta Books, 150 F. Supp. 2d 613, 614–17 (S.D.N.Y. 2001), *aff'd*, 283 F.3d 490 (2d Cir. 2002).
[138] *Id.* at 620–22 (quoting Boosey & Hawkes Music Publishers, Ltd v. Walt Disney Co., 145 F.3d 481, 487 n.3 (2d Cir. 1998)).
[139] *See supra* Section II(A)(1)–(2).
[140] *Random House*, 150 F. Supp. 2d at 622–623.
[141] Complaint, Class Action, & Demand for Jury Trial at 13-18, Doe v. GitHub, Inc., No. 22-cv-06823 (N.D. Cal. Nov. 3, 2022).
[142] *Id.* at 12.
[143] *Id.*

licenses that require attribution, copyright notices, and/or inclusion of certain terms in subsequent licenses, but none of these license provisions are observed in the generated code output by the generative AI systems.[144]

The *Doe* plaintiffs rely on claims other than copyright infringement, such as DMCA and breach-of-contract claims, presumably to avoid the fair use affirmative defense, but also because the defendants potentially had a license to use the software code for training the generative AI systems through the GitHub terms of service.[145] Specifically, the GitHub terms of service contain a license granted by all users to GitHub to "'store, archive, parse, and display . . . and make incidental copies' as well as 'parse it into a search index or otherwise analyze it' and 'share' the content in public repositories with other users."[146] The defendants assert that these uses encompass the use of the content for training the generative AI systems,[147] but the facts of training and implementing a generative AI system may be outside the scope of the expressly licensed uses. While seemingly broad grants in a license agreement may appear to cover the use of copyrighted works in generative AI systems, there is a significant chance that a court would find such use a "new use" because the right to incorporate a copyrighted work into a generative AI system may be "so dissimilar" from existing technological uses of the work "as to preclude consideration of [generative AI] rights as even falling within the 'ambiguous penumbra' of the terms used in the agreement."[148]

The Supreme Court decision in *Andy Warhol Foundation v. Goldsmith* provides instructive guidance on use of copyrighted works by

---

[144] *Id.* at 21, 39.

[145] *Id.* at 22 ("The Fair Use affirmative defense is only applicable to Section 501 copyright infringement. It is not a defense to violations of the DMCA, breach of contract, nor any other claim alleged herein. It cannot be used to avoid liability here."); *see* John A. Rothchild & Daniel H. Rothchild, *Copyright Implications of the Use of Code Repositories to Train a Machine Learning Model*, FREE SOFTWARE FOUND. (Feb. 24, 2022, 6:23 PM), https://www.fsf.org/licensing/copilot/copyright-implications-of-the-use-of-code-repositories-to-train-a-machine-learning-model (discussing possibility that express grants of rights from users to GitHub in GitHub terms of service cover use of user-uploaded content for AI training).

[146] Defendants GitHub and Microsoft's Notice of Motions and Motions to Dismiss Operative Complaint in Consolidated Actions at 4–5, No. 22-cv-06823 (N.D. Cal. Jan. 26, 2023) ("Anyone is free to examine, learn from, and understand that code, as well as repurpose it in various ways. And, consistent with this open source ethic, neither GitHub's TOS nor any of the common open source licenses prohibit either humans or computers from reading and learning from publicly available code.").

[147] *Id.*

[148] Random House, Inc. v. Rosetta Books LLC, 150 F.Supp.2d 613, 623 (S.D.N.Y. 2001) (quoting Tele–Pac, Inc. v. Grainger, 168 A.D.2d 11 (1st Dep't 1991)).

a licensed party when such use exceeds the scope of the license. In *Andy Warhol Foundation v. Goldsmith*, the photographer Goldsmith granted the magazine Vanity Fair a license permitting the magazine to publish an illustration by the artist Andy Warhol based on Goldsmith's photograph of the musical artist Prince. While Warhol in a sense exceeded the scope of the license by preparing additional work based on the photograph, the Second Circuit noted that "[o]f course, if a secondary work is sufficiently transformative, the fact that its 'raw material' was acquired by means of a limited license will not necessarily defeat a defense of fair use."[149] While the Court's decision did not turn on the scope of Goldsmith's license being exceeded, the Court found that both the Andy Warhol Foundation's licensing of the Warhol illustrations and Goldsmith's photograph "share substantially the same purpose," which when considered with the commercial nature of the license, "counsel against fair use" under the "purpose and character of the use" factor of the fair use analysis. [150] As a result, the scope of a license grant for both the original work and the transformed work may contribute to how a court analyzes a fair use defense.[151]

Another consideration for the scope of the license is whether the license includes a grant to sublicense. If a license for a copyrighted work does not expressly grant a right to sublicense, then the licensee cannot sublicense the copyrighted work.[152] There is no implied right to sublicense in an implied license.[153] However, some courts might allow an implied sublicense to arise between sublicensee and a licensee with an expressly granted right to sublicense from the original licensor.[154]

---

[149]11 F.4th 26, 34, 50 n.13 (2d Cir. 2021), *cert. granted*, 142 S. Ct. 1412 (2022); *see also* Petition for Certiorari, Andy Warhol Found. for Visual Arts, Inc. v. Goldsmith, No. 21-869 (Dec. 9, 2021).

[150] Andy Warhol Found. for Visual Arts, Inc. v. Goldsmith, 598 U.S. 25 (2023); 17 U.S.C. § 107(1).

[151] *E.g.*, *id.* at 33 ("To hold otherwise would potentially authorize a range of commercial copying of photographs, to be used for purposes that are substantially the same as those of the originals. As long as the user somehow portrays the subject of the photograph differently, he could make modest alterations to the original, sell it to an outlet to accompany a story about the subject, and claim transformative use.").

[152] Gardner v. Nike, Inc., 279 F.3d 774, 780 (9th Cir. 2002) ("[T]he 1976 Act does not allow a copyright licensee to transfer its rights under an exclusive license, without the consent of the original licensor.").

[153] Catalogue Creatives, Inc. v. Pac. Spirit Corp. No. CV 03-966-MO, 2005 WL 1950231, at *2 (D. Or. Aug. 15, 2005) (holding that "the recipient of an implied license may [not] grant an implied sublicense by its conduct").

[154] Photographic Illustrators Corp. v. Orgill, Inc., 953 F.3d 56, 64 (1st Cir. 2020) (holding that "where a licensor grants to a licensee the unrestricted right to sublicense and permit others to use a copyrighted work, a sublicense may be implied by the conduct of the sublicensor and sublicensee"); *see also* Rohena Rajbhandari, Comment,

Whether there is a valid sublicense within the scope of the license agreements may determine the validity of some conduct, such as dataset assembly via web scraping.[155]

3. Implied Agreements

While express agreements were discussed extensively in Part II, there is a possibility that implied license doctrine may be important for evaluating the relationship between certain parties, such as between an Internet platform and a dataset assembler, especially when the interaction of the dataset assembler with the Internet platform may be automated and not involve accessing the Internet platform in a way that signals assent to the platform's terms of service.

For example, in *Field v. Google Inc.*, the court interpreted a website operator's decision to omit "no-archive" meta-tags that would avoid caching of the webpage as conduct amounting to permission to cache the webpage.[156] Specifically, the court found an implied license for the conduct that overcame a claim of copyright infringement.[157] While this might suggest some flexibility for dataset assemblers who scrape websites to collect content for AI training datasets, the *Field* court noted that Field "was aware of these industry standard mechanisms" for web crawling and thus had "knowledge of how Google would use the copyrighted works he placed on those pages, and . . . knowledge that he could prevent such use."[158] Thus, the finding of an implied license was tied to the plaintiff's awareness of search engine caching.[159] In contrast, most copyright owners with their works hosted on the Internet may not be aware of how dataset assemblers are scraping the web to create training datasets, or that their content may be used to train generative AI systems. Further, if a court found the use of copyrighted works for training generative AI systems to not be fair in a particular instance, the court might also decline to find an implied license to use that copyrighted work for that use.

As an illustrative example, in *Associated Press v. Meltwater U.S.*

License to Sublicense: The Legal Possibility of Impliedly Sublicensing a Copyrighted Work, 62 B.C. L. REV. E. Supp. Il-425, Il.-432, Il.-435 (2021).

[155] *See supra* text accompanying note 78-79.

[156] Field v. Google Inc., 412 F. Supp. 2d 1106, 1116 (D. Nev. 2006).

[157] *Id.* at 1116 ("Consent to use the copyrighted work need not be manifested verbally and may be inferred based on silence where the copyright holder knows of the use and encourages it.").

[158] *Id.*

[159] *Id.*

Holdings, Inc., Meltwater News scraped the Internet, much "[l]ike Internet search engines," to collect news articles and create an index of the content.[160] Meltwater used this index to provide search functionality for news articles, provide excerpts of the news articles, and allow subscribers to "archive any of their search results in a personal archive stored on Meltwater's database."[161] The court held that Meltwater's use of Associated Press's news articles in this way was not fair use because there was "nothing transformative" about Meltwater's use which amounted to "a classic news clipping service."[162]

As another affirmative defense, Meltwater argued that there was an implied license to copy and distribute the copyrighted works.[163] However, in the absence of any evidence that there was a "meeting of the minds" between Meltwater and AP or that any of the implied-license factors were satisfied, the court rejected Meltwater's argument that the lack of a robots.txt protocol on the website forbidding the automated crawling and scraping of the website gave rise to an implied license.[164] Thus, dataset assemblers and AI creators should not rely on an implied license to reproduce and distribute copyrighted works as a defense for instances where their use of the copyrighted works exceeds any express agreements or where such express agreements do not exist, because a court finding an implied license is unlikely if the same court does not find fair use.[165]

4. Implied Agreements

Even if the terms of service for a website hosting copyrighted works included provisions forbidding the copying of works, there is currently an unresolved circuit split regarding whether the Copyright

---

[160] 931 F. Supp. 2d 537, 544 (S.D.N.Y. 2013).
[161] *Id.* at 544–46.
[162] *Id.* at 556, 561 ("Meltwater's business model relies on the systematic copying of protected expression and the sale of collections of those copies in reports that compete directly with the copyright owner and that owner's licensees and that deprive that owner of a stream of income to which it is entitled.").
[163] *Id.* at 561–62 ("The test for determining whether an implied license exists in the copyright context has three elements. The defendant must show that (1) the licensee requested the creation of a work; (2) the licensor made that particular work and delivered it to the licensee who requested it; and (3) the licensor intended that the licensee copy and distribute his work.").
[164] *Id.* at 562–64.
[165] *See id.* at 564 ("It is worth observing that, when a crawler is making a fair use of a website's content, it does not need to resort to the implied license doctrine; where it does not [make fair use of the website's content], then the website's failure to use the robots.txt protocol to block its access will not create an implied license.").

Act preempts breach-of-contract claims.[166] In order to achieve a uniform body of federal copyright law, section 301(a) of the Copyright Act of 1976 expressly preempts "all legal or equitable rights that are equivalent to any of the exclusive rights" of the Copyright Act.[167] After the Seventh Circuit held in *ProCD, Inc. v. Zeidenberg* that section 301 "does not itself interfere with private transactions in intellectual property,"[168] legal scholars and courts alike were unsure where the boundaries of section 301 were and whether private parties could contract around federal statutory rights like fair use.[169] The question of preemption is especially important in light of the widespread use of boilerplate provisions.[170]

One contemporary case that might help clarify whether section 301 preempts breach-of-contract claims and provides guidance to lower courts is currently pending a decision by the Supreme Court on whether to certify a petition for certiorari.[171] In *ML Genius Holdings LLC v. Google LLC*, two Internet platforms that provide song lyric transcriptions, Genius and LyricFind, both obtained licenses from music publishers for the right to display and distribute lyrics.[172] Genius used "digital watermarks" to reveal that LyricFind was copying song lyric transcriptions hosted by Genius and licensing these copies to Google.[173] Genius then sued Google and LyricFind for breaching

---

[166] *See* Jaci L. Overmann, *With End-User License Agreements, Which Will Prevail: Copyright Rights or Contract Rights?*, NAT'L L. REV. (Dec. 2, 2022), https://www.natlawreview.com/article/end-user-license-agreements-which-will-prevail-copyright-rights-or-contract-rights.

[167] 17 U.S.C. § 301(a).

[168] ProCD, Inc. v. Zeidenberg, 86 F.3d 1447, 1455 (7th Cir. 1996) ("Terms and conditions offered by contract reflect private ordering, essential to the efficient functioning of markets.").

[169] *See, e.g.*, Guy A. Rub, *Copyright Survives: Rethinking the Copyright-Contract Conflict*, 103 VA. L. REV. 1141, 1146 (2017) (discussing the no-preemption approach and the facts-specific approach to the preemption of contract claims by copyright law); Viva R. Moffat, *Super-Copyright: Contracts, Preemption, and the Structure of Copyright Policymaking*, 41 U.C. DAVIS L. REV. 45 (2007) (arguing that courts should not allow contract provisions to supersede the federal copyright system); Guy A. Rub, *Against Copyright Customization*, 107 IOWA L. REV. 677 (2022) (discussing the courts' inability to find a workable balance between contract and copyright law).

[170] Rub, *supra* note 171, at 683 ("With those boilerplate agreements, corporations can summon copyright law's powerful enforcement mechanisms at will to crush activities that copyright law is not supposed to prohibit.").

[171] Petition for Certiorari at i, *ML Genius Holdings LLC v. Google LLC*, Sup. Ct. No. 22-121 (U.S. Aug. 5, 2022), *cert. denied*, June 26, 2023.

[172] Genius Media Grp. Inc. v. Google LLC, No. 19-CV-7279 (MKB), 2020 WL 5553639, at *1 (E.D.N.Y. Aug. 10, 2020), *aff'd sub nom.* ML Genius Holdings LLC v. Google LLC, No. 20-3113, 2022 WL 710744 (2d Cir. Mar. 10, 2022).

[173] *ML Genius Holdings LLC*, 2022 WL 710744, at *22a.

Genius's terms of service.[174] The district court found and the Second Circuit affirmed that section 301(a) preempted Genius's contract claim because the "breach of contract claim is not qualitatively different from a copyright claim."[175] Specifically, Genius's terms of service conditioned access to lyrics hosted on the Genius website with a promise not to "copy, modify, sell and/or distribute content appearing on Genius's website," which the court found to be equivalent to the exclusive rights to reproduce, prepare derivative works, distribute copies provided by section 106 of the Copyright Act.[176]

If an Internet platform is hosting copyrighted works and is licensed to do so, similar to how Genius licenses the copyrights to the lyrics that it hosts, the platform itself may be restrained from using contract law to confront entities who scrape the content from the platform. This might prove doubly difficult if those same entities are also licensing the content from the copyright owners. Additional guidance from the Supreme Court may be helpful for identifying how to properly balance contract interests with copyright interests, though the preemption issue may not be particularly helpful for copyright owners.[177]

B. Proposals For Mitigating Infringement Liability

Although questions regarding fair use and the interface of contracts with copyrights vis-à-vis generative AI systems may be in flux as the technology rapidly evolves and the law catches up, there are various considerations that could help mitigate the liability of generative AI stakeholders should the fair use defense fail to insulate them from claims of copyright infringement.

1. Implement Copyright-Protective Guardrails

Importantly, dataset assemblers and AI creators should implement copyright-protective mechanisms or "guardrails" in their procedures and design structures that could at least insulate them from meeting or satisfying the volition requirement for direct

---

[174] *Id.* at *4a.
[175] *Id.* at *12a-13a.
[176] *Id.* at *11a; 17 U.S.C. § 106.
[177] *See* Elazari Bar On, *supra* note 48, at 590 ("[T]raditional solutions such as preemption and misuse are specifically ill-equipped to address problems created by adherent-creator types of IP boilerplate.").

infringement.[178] For example, dataset assemblers should consider how to collect content along with suitable rights to such content, such that they can make representations and warranties regarding the datasets that they provide. Generative AI systems could be designed to accommodate upstream licensing requirements, such as attribution and copyright notices.[179] They might also be designed to include technical measures that would prevent outputting content that is substantially similar to a work used for training.[180] Furthermore, AI creators who use license agreements that specify substantial use restrictions on the generative AI systems, such as the RAILs, should diligently enforce these license agreements in order to assuage public fears of an unhinged, illegal, irresponsible AI.[181] Implementing strong copyright-protective guardrails could strengthen the notion that generative AI systems have substantial non-infringing uses, thereby minimizing indirect infringement liability for AI creators and dataset assemblers.[182]

2. Improve License Agreements

Ideally, license agreements would resolve many of the infringement liability issues that arise from generative AI systems, but the extreme breadth of data used for training generative AI systems means that parties rely on the generic terms of service that are ubiquitous on the Internet for any such licenses, and these terms-of-service licenses do not provide the strongest shield from liability.[183] Nevertheless, license agreements could be improved and updated to account for the scope of the use, such as for training and using generative AI systems.[184] Further, license agreements should seek to balance the distribution of power between parties by giving more deference to copyright owners, thereby

---

[178] *See* VHT, Inc. v. Zillow Grp., Inc., 918 F.3d 723, 733 (9th Cir. 2109).
[179] *See* Doe v. GitHub, Inc. Complaint, *supra* note 141, at 21 ("Codex and Copilot were not programmed to treat attribution, copyright notices, and license terms as legally essential. Defendants made a deliberate choice to expedite the release of Copilot rather than ensure it would not provide unlawful Output.").
[180] Kyle Wiggers, *The Current Legal Cases Against Generative AI are Just the Beginning*, TECHCRUNCH (Jan. 27, 2023, 8:30 AM), https://techcrunch.com/2023/01/27/the-current-legal-cases-against-generative-ai-are-just-the-beginning/ ("For Copilot, GitHub introduced a filter that checks code suggestions with their surrounding code of about 150 characters against public GitHub code and hides suggestions if there's a match or 'near match.' It's an imperfect measure — enabling the filter can cause Copilot to omit key pieces of attribution and license text").
[181] *See supra* text accompanying notes 109-112.
[182] Sony Corp. of Am. v. Universal City Studios, Inc., 464 U.S. at 442 (1984).
[183] *See supra* Section III(A).
[184] *See supra* Section III(A)(2).

strengthening the scope of the licenses while lessening any potential unconscionability.[185]

3. Address the High Transaction Costs for Copyright Clearance

Ideally, care should be taken to ensure that appropriate rights are cleared so that copyrighted works may be used for generative AI systems. However, given the enormous number of works in a given training dataset, the transaction costs for fully clearing a single dataset would be unmanageable. To balance the interest of copyright owners in controlling whether their works are used in generative AI systems, AI creators could offer an opt-out or opt-in system.[186] However, an opt-out or opt-in system may not be effective or desirable from the perspective of copyright owners.[187] Alternatively, Congress could enact legislation that allows the use of copyrighted works for training, and establishes a system for compensating copyright owners. Under such a system, organizations that collect compulsory license fees for works included in training datasets and distribute the fees to copyright owners.[188] This approach has the potential to balance the property interests of copyright owners with the general public interest in the possibilities of generative AI systems, but care must be taken to ensure this balance.[189]

CONCLUSION

As generative AI systems continue to advance, the legal

---

[185] *See supra* text accompanying notes 123-124.

[186] Stability AI provides a tool allowing copyright owners to determine whether their images are in the LAION datasets and to opt out of their use in training generative AI systems. *See* HAVE I BEEN TRAINED?, https://haveibeentrained.com/ (last visited Feb. 19, 2023).

[187] In *Authors Guild v. Google, Inc.*, the parties reached an "opt-out" settlement agreement that would have allowed copyright owners to opt out of their works being digitized by Google, but the District Court rejected this system due to copyright concerns and urged the parties to adopt an opt-in system instead. 770 F. Supp. 2d 666, 680–82, 686 (S.D.N.Y. 2011); *see also* Alison Flood, *Thousands of Authors Opt Out of Google Book Settlement*, THE GUARDIAN (Feb. 23, 2010), https://www.theguardian.com/books/2010/feb/23/authors-opt-out-google-book-settlement.

[188] Lemley & Casey, *supra* note 6 at 759. Congress has already created compulsory license systems for specific industries such as music and television rebroadcasting. *See* 17 U.S.C. §§ 111, 114–115, 119.

[189] *See, e.g.*, Jacob Victor, *Reconceptualizing Compulsory Copyright Licenses*, 72 STAN. L. REV. 915 (2022) (discussing the imbalance between the competing interests of incentives and access that arises when royalty-rate setting is influenced by free market policy rather than copyright policy).

relationship between copyright owners, dataset assemblers, and AI creators will also grow more complex. While intellectual property boilerplate provisions and adhesive contracts currently dominate the contractual relationships between these parties, additional care should be taken to improve license agreements between parties and to balance the distribution of power. In this way, the AI landscape may continue to rapidly evolve and flourish while humans may feel empowered and incentivized to continue to create new, original works.