DOES TRAINING AI VIOLATE COPYRIGHT LAW?

Jenny Quang[†]

I. INTRODUCTION

From targeted advertising to search optimization, machine learning (ML) algorithms are increasingly prevalent in our daily lives.[1] Artificial intelligence (AI) and copyright law intersect when copyrighted data are used to train machines to learn, reason, and act as humans do. The development of some AI technologies, such as autonomous vehicle software, facial recognition algorithms, and smart assistants, requires massive data mining of video, photos, or text that may be subject to copyright.[2] However, the downloading and storage of copyrighted data to train machine learning models may violate copyright law and impose undue liability on AI developers. This Note identifies how Congress can drive AI innovation by clarifying copyright law and adopting a safe harbor for data mining.

Recently, a 3-D imaging firm filed copyright claims accusing Meta and Princeton University of illegally downloading its data for use in scene-recognition AI projects.[3] Although it is uncertain how the court will rule on this novel issue, it is clear that the stakes are high—the market for scene-recognition technology is estimated to reach $60 billion by 2025, and the cost of the data collection at issue is in the millions of dollars.[4] Alongside potentially steep damages, the plaintiffs are seeking injunctive relief from further acts of copyright infringement and destruction of all infringing copies.[5] The outcome of this case, and others like it that may follow, could handicap new AI technologies.

Most copyright stakeholders believe that the fair use doctrine is sufficient to defend data mining uses in the United States.[6] However, fair use is an undesirable form of protection because the doctrine has been stretched beyond

DOI: https://doi.org/10.15779/Z38XW47X3K

[†] J.D., 2022, University of California, Berkeley, School of Law.

[1] *See* ANDREW NG, MACHINE LEARNING YEARNING 6 (2018).

[2] *See infra* Part II.

[3] *See* Complaint for Copyright Infringement at 27–30, UAB "Planner5D" v. Facebook, Inc., No. 3:20-cv-08261 (N.D. Cal. Nov. 23, 2020).

[4] *Id.* at 3.

[5] *Id.* at 30.

[6] U.S. PAT. & TRADEMARK OFF., PUBLIC VIEWS ON ARTIFICIAL INTELLIGENCE AND INTELLECTUAL PROPERTY POLICY 26 (2020) ("Most commenters found that existing law does

its limits and is often unpredictable in practice. This uncertainty disproportionately handicaps smaller actors; the possibility of copyright liability may intimidate new AI developers aiming to compete with established tech giants and could influence how researchers conduct their work. In a recent example, an AI-based legal research startup shut down amid financial pressures brought on by a copyright infringement lawsuit.[7] And statutory fines, which range from $200 to $150,000 per work,[8] are unnecessarily crippling, especially when a single machine learning model may be trained using thousands to millions of works.[9] Establishing a clear right to use copyrighted materials for data mining is consistent with the goals of copyright law and would remove barriers to innovation.

A safe harbor for data mining is warranted because the use of data to develop functional AI technologies is *fundamentally* not an act of infringement. The seminal Supreme Court copyright case *Baker v. Selden* distinguished a copyrighted work from its material form and showed that not all uses of a work's material form are acts of copyright infringement.[10] Copyright infringement requires not just copying of a work's material form but also the unauthorized use of the work for its expressive purpose. Merely technical or non-communicative uses are not uses of a work for its expressive purpose and therefore are not copyright infringement.

not require modification, as fair use is a flexible doctrine and is capable of adapting to the use of copyrighted works in an AI context.").

[7] Lyle Moran, *ROSS Intelligence Will Shut Down Amid Lawsuit from Thomson Reuters*, ABA J. (Dec. 11, 2020, 11:50 AM), https://www.abajournal.com/news/article/ross-intelligence-to-shut-down-amid-thomson-reuters-lawsuit ("Litigation is expensive—no matter how speculative the claims against you nor how worthy your position[.] . . . With our company ensnared by this legal battle, we have been unable to raise another round of funding to fuel our development and marketing efforts. Our bank account is running out, and we must cease operations in the new year." (quoting ROSS Intelligence)).

[8] 17 U.S.C. § 504.

[9] *See, e.g.*, Tom B. Brown et al., Language Models are Few-Shot Learners 1 (July 22, 2020) (unpublished manuscript), https://arxiv.org/pdf/2005.14165.pdf ("[T]his method still requires task-specific fine-tuning datasets of thousands or tens of thousands of examples."); Chen Sun, Abhinav Shrivastava, Saurabh Singh & Abhinav Gupta, Revisiting Unreasonable Effectiveness of Data in Deep Learning Era 3 (Aug. 4, 2017) (unpublished manuscript), https://arxiv.org/pdf/1707.02968.pdf (training on a dataset of "300M images and 375M labels").

[10] *See* Baker v. Selden, 101 U.S. 99 (1879); ABRAHAM DRASSINOWER, WHAT'S WRONG WITH COPYING? 88-100 (2015). In *Baker*, the copyrighted work was a book that explained a novel accounting method, and its material form included the accounting forms used as part of the explanation. *See* 101 U.S. at 100–01. The defendant's non-communicative copying of the accounting forms to perform the novel accounting method did not constitute infringement. *See id.* at 107.

Likewise, to download copyrighted images and text for data mining is to make copies for a different purpose. Training a machine learning model with this copyrighted data does not infringe because the data are not redistributed or recommunicated to the public. Copyright protects creative expression, but model training extracts unprotectable ideas and patterns from data. Thus, data mining uses of copyrighted works need not even be subject to a fair use analysis.

Copyright law distinguishes between creative expression and unprotectable ideas.[11] In this Note, "data mining" will specifically refer to the mining of *expressive* data (i.e., literary works, photographs, video) for *functional*, or non-expressive, purposes. Expressive applications of data mining (i.e., AI-generated art, music, and literature) are outside of the scope of this analysis. Due to an increasing interest in artificial intelligence technologies,[12] this Note will focus on copyright and data mining in an artificial intelligence and machine learning context.

This Note argues that Congress should adopt a safe harbor for data mining of copyrighted works because (1) it is fundamentally not an act of infringement and (2) legal uncertainty currently exists with the fair use doctrine. The Note presents the argument as follows. Part II explains how data mining implicates copyright law and reviews how courts have applied the fair use doctrine unevenly in similar technological contexts. Next, Part III argues that data mining is fundamentally not copyright infringement and that a safe harbor is needed to clear up legal uncertainty that results from the fair use doctrine. Finally, Part IV reviews data mining exceptions in foreign jurisdictions and outlines a legislative proposal for a safe harbor in U.S. copyright law.

II. BACKGROUND

A. DATA MINING DEFINITIONS

Broadly, data mining involves improving future decisions by finding patterns in data collected from past events.[13] Interest in the field has rapidly grown with advancements in data collection and storage, the use of machine learning to process this data, and the falling cost of computational power.[14]

[11] *See* 17 U.S.C. § 102(b).

[12] *See, e.g.*, Exec. Order No. 13,859, 84 Fed. Reg. 3967 (Feb. 14, 2019).

[13] *See* Tom M. Mitchell, *Machine Learning and Data Mining*, 42 COMMC'NS ACM 30, 30 (1999).

[14] *See id.*

Machine learning is a process of "[u]sing data to answer questions."[15] Data is key to this process, as machine learning algorithms use statistics to find insights hidden in massive amounts of data.[16] These algorithms are also responsible for the majority of recent advancements and applications in artificial intelligence, a field that trains machines to learn, reason, and act as humans do.[17]

Two subfields of AI learn from expressive data and will be used as examples throughout this Note. Computer vision is a subfield of AI that studies visual data.[18] A computer is taught how to understand the world through images and video.[19] Applications of such technology include autonomous vehicles, facial recognition, and medical imaging.[20] Natural language processing (NLP) uses computational techniques to understand and represent human languages.[21] Large datasets of text are required to train models to understand and generate new text.[22] Applications include translation, information extraction, and question answering.[23]

The majority of artificial intelligence applications require machine learning algorithms and models.[24] A machine learning algorithm will find patterns from past data and output a machine learning model that captures these patterns.[25] The model can then be used to make predictions on new data.[26] After first defining the problem (i.e., what the model should predict), an AI developer must collect data, prepare the data (including labeling or annotating), choose a model architecture; train the model with the annotated data ("training data")

[15] Yufeng G, *What is Machine Learning?*, TOWARDS DATA SCI. (Aug. 24, 2017), https://towardsdatascience.com/what-is-machine-learning-8c6871016736.

[16] *Id.*

[17] Karen Hao, *What is Machine Learning?*, MIT TECH. REV. (Nov. 17, 2018), https://www.technologyreview.com/2018/11/17/103781/what-is-machine-learning-we-drew-you-another-flowchart; Karen Hao, *What is AI? We Drew You a Flowchart to Work it Out*, MIT TECH. REV. (Nov. 10, 2018), https://www.technologyreview.com/2018/11/10/139137/is-this-ai-we-drew-you-a-flowchart-to-work-it-out.

[18] Ben Dickson, *What is Computer Vision?*, PCMAG (Feb. 9, 2020), https://www.pcmag.com/news/what-is-computer-vision.

[19] *Id.*

[20] *Id.*

[21] K. R. CHOWDHARY, FUNDAMENTALS OF ARTIFICIAL INTELLIGENCE 604 (2020).

[22] *Id.* at 608.

[23] *Id.* at 608–09.

[24] *See* Hao, *What is AI?*, *supra* note 17.

[25] *Training ML Models*, AMAZON WEB SERVS., https://docs.aws.amazon.com/machine-learning/latest/dg/training-ml-models.html.

[26] *Id.*

using a machine learning algorithm, and test the model with new data ("testing data").[27]

In practice, AI developers will often fine-tune open source models that have been pre-trained on extremely large datasets.[28] For example, a developer who wants to train a computer to recognize the movements of a surgeon's hand during an operation will start with a model that has already been trained using ImageNet, a database containing over 14 million images of common objects (i.e., cat, dog, car).[29] This pre-trained model will already be able to find edges and shapes and recognize common objects.[30] The developer will then compile her own dataset of surgical images (containing hundreds to thousands of photographs) and fine-tune the model on that training data. The new model would then be able to better recognize surgery-specific objects, such as surgical tools and gloved hands. This fine-tuning ("transfer learning") process contributes to the democratization of AI technology by allowing smaller innovators to more easily develop their own products when they build off previous work.[31]

However, the threat of copyright liability has the potential to stifle this democratization process. Starting with a pre-trained model can reduce the data points needed from millions to thousands.[32] But an AI developer will still need to download some data to fine-tune her model. Because machine learning

[27] *See id.*; Zaid Alissa Almaliki, *Do You Know How to Choose the Right Machine Learning Algorithm Among 7 Different Types?*, TOWARDS DATA SCI. (Mar. 19, 2019), https://towardsdatascience.com/do-you-know-how-to-choose-the-right-machine-learning-algorithm-among-7-different-types-295d0b0c7f60; Yufeng G, *The 7 Steps of Machine Learning*, TOWARDS DATA SCI. (Aug. 31, 2017), https://towardsdatascience.com/the-7-steps-of-machine-learning-2877d7e5548e.

[28] *See* Naveen Joshi, *How to Fine-tune Your Artificial Intelligence Algorithms*, ALLERIN (Jan. 13, 2020), https://www.allerin.com/blog/how-to-fine-tune-your-artificial-intelligence-algorithms; Aidan Boyd, Adam Czajka & Kevin Bowyer, *Deep Learning-Based Feature Extraction in Iris Recognition: Use Existing Models, Fine-tune or Train From Scratch*?, 10 IEEE INT'L CONF. ON BIOMETRICS THEORY, APPLICATIONS AND SYS., SEPT. 2019, at 1.

[29] *Summary and Statistics*, IMAGENET (Apr. 30, 2010), http://image-net.org/about-stats.

[30] *See* Sebastian Ruder, *NLP's ImageNet Moment Has Arrived*, THE GRADIENT (July 8, 2018), https://thegradient.pub/nlp-imagenet ("Importantly, knowledge of edges, structures, and the visual composition of objects is relevant for many CV tasks . . . . A key property of an ImageNet-like dataset is thus to encourage a model to learn features that will likely generalize to new tasks in the problem domain.").

[31] *See, e.g.*, Vikash Gupta et al., Democratizing Artificial Intelligence in Healthcare: A Study of Model Development Across Two Institutions Incorporating Transfer Learning. 1 (Sept. 25, 2020) (unpublished manuscript), https://arxiv.org/ftp/arxiv/papers/2009/2009.12437.pdf ("[I]f a well-tested deep learning model from another institution is available, it can be adopted for use in model fine-tuning on a relatively smaller local dataset, thereby allowing institutions with fewer resources to also directly participate in AI development.").

[32] *See* Brown et al., *supra* note 9, at 1 ("[T]his method still requires task-specific fine-tuning datasets of thousands or tens of thousands of examples.").

models are trained on massive amounts of data, this step is often automated.[33] In computer vision, images and videos may be scraped from the internet (e.g., Google Images, YouTube) and incorporated into large datasets for training.[34] In NLP, corpora of text are usually mined from the internet (e.g., Wikipedia, Gmail).[35] The downloading of data that may be subject to copyright presents a potential violation of copyright law.

B. COPYRIGHT AND DATA MINING

The Intellectual Property Clause of the U.S. Constitution grants Congress the power to "promote the Progress of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries."[36] Copyright protection serves a dual purpose: to motivate the creative activity of authors and artists and to advance public welfare through access to expressive works.[37] In exchange for the creation of literary and artistic works, authors and artists are granted certain exclusive rights, including the rights to reproduction, distribution, and preparation of derivative works, for a limited period.[38] To enforce these rights, copyright owners can sue alleged infringers for monetary relief, including statutory damages, which can range from $200 to $150,000 per work depending on the willfulness of infringement.[39] These steep fines provide an incentive for copyright owners to enforce copyright interests when harm from infringement can be difficult to prove.[40]

[33] *See* Lewis Chou, *Four Basic Ways to Automate Data Extraction*, TOWARDS DATA SCI. (Oct. 12, 2019), https://towardsdatascience.com/four-basic-ways-to-automate-data-extraction-3151064dc110.

[34] *See, e.g.*, Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar & Li Fei-Fei, *Large-scale Video Classification with Convolutional Neural Networks*, 2014 PROC. IEEE CONF. ON COMPUT. VISION & PATTERN RECOGNITION 1725 (using dataset of 1 million YouTube videos); Adrian Rosebrock, *How to Create a Deep Learning Dataset Using Google Images*, PYIMAGESEARCH (Dec. 4, 2017), https://www.pyimagesearch.com/2017/12/04/how-to-create-a-deep-learning-dataset-using-google-images.

[35] *See, e.g.*, Evgeniy Gabrilovich & Shaul Markovitch, *Wikipedia-based Semantic Interpretation for Natural Language Processing*, 34 J. A.I. RSCH. 443 (2009); Mia Xu Chen et al., Gmail Smart Compose: Real-Time Assisted Writing 3 (May 17, 2019) (unpublished manuscript), https://arxiv.org/pdf/1906.00080.pdf (using dataset of "user-composed e-mails").

[36] U.S. CONST. art. I, § 8, cl. 8.

[37] *See* 1 MELVILLE B. NIMMER & DAVID NIMMER, NIMMER ON COPYRIGHT § 1.03 (2021).

[38] 17 U.S.C. § 106.

[39] 17 U.S.C. § 504.

[40] Peter S. Menell, *This American Copyright Life: Reflections on Re-equilibrating Copyright for the Internet Age*, 61 J. COPYRIGHT SOC'Y U.S.A. 235, 306–07 (2014).

Artificial intelligence and copyright law intersect when expressive data is used to train machines to learn, reason, and act as humans do. Under § 106, the reproduction right grants a copyright holder the exclusive right to make copies of the protected work.[41] As explained in Section II.A, developers often use images, video, and text downloaded from the internet to train machine learning models.[42] The downloaded data are essentially copies that are stored via hard drives, cloud storage, or other data repositories.[43] Given the large volume of data—often scraped from the internet en masse—that is needed to train a machine learning model, it is likely that some of that training data is protected by copyright.[44] Because copyright infringement is a strict liability offense,[45] it does not matter if a developer was unaware that copyrighted works existed in the dataset.

The download and storage of data creates copies that may be subject to copyright. The Copyright Act states that copies are "material objects . . . in which a work is fixed *by any method now known or later developed*, and from which the work can be perceived, reproduced, or otherwise communicated, either directly or *with the aid of a machine or device*."[46] Thus, data that are solely used by machines during model training would be treated just the same as works perceived by humans under copyright law. Further, a work is "fixed" in a tangible medium when its embodiment in a copy is "*sufficiently permanent or stable* . . . for a period of more than transitory duration."[47]

Copyright caselaw dealing with memory storage and RAM shows that training data downloaded onto a computer's hard drive would be sufficiently fixed under copyright law. In *Stern Electronics, Inc. v. Kaufman*, the Second Circuit held that audiovisual works of a video game that are permanently embodied in "memory devices" of the game were protectable under copyright law.[48] Later, in *MAI Systems Corp. v. Peak Computer, Inc.*, the Ninth Circuit held that loading

[41] 17 U.S.C. § 106.

[42] *See supra* Section II.A.

[43] *See* Jim Dowling, *Guide to File Formats for Machine Learning: Columnar, Training, Inferencing, and the Feature Store*, TOWARDS DATA SCI. (Oct. 25, 2019), https://towardsdatascience.com/guide-to-file-formats-for-machine-learning-columnar-training-inferencing-and-the-feature-store-2e0c3d18d4f9.

[44] Contemporary photographs are presumptively under copyright. Brammer v. Violent Hues Prods., LLC, 922 F.3d 255, 266 (4th Cir. 2019).

[45] 17 U.S.C. § 106; *see, e.g.*, *Brammer*, 922 F.3d at 265 ("As a basic matter, copyright infringement is a strict liability offense, in which a violation does not require a culpable state of mind.").

[46] 17 U.S.C. § 101 (emphases added).

[47] *Id.* (emphasis added).

[48] 669 F.2d 852, 855–56 (2d Cir. 1982).

of copyrighted software into a computer's random-access memory (RAM) created an unauthorized reproduction under the Copyright Act.[49] Thus, fixation of a copy in either permanent or temporary memory storage is sufficiently permanent and stable to satisfy the Copyright Act.

So, even if training data is not to be *permanently* stored on a hard drive, like the audiovisual works in *Stern Electronics* (a developer could, in theory, delete the training data from her computer once she finished training her model), the storage could still constitute unauthorized copying because a hard drive is more stable than RAM.[50] And what about training data that is not stored on a personal hard drive, but on a cloud storage service?[51] Technically, cloud storage is still the storage of digital data on physical hard drives, but in off-site locations that are accessed via the internet or a private network connection.[52] The Copyright Act also accounts for methods of fixation "now known or later developed," which would allow for extrapolation to future data storage devices, like solid-state drives.[53] Thus, whether training data is downloaded onto a personal computer or on the cloud, the copies should be sufficiently fixed in a medium to implicate copyright law.

An allegedly infringing developer could argue that although she downloaded copyrighted images to train her computer vision model, there is no trace of the copyrighted works in her final model. If a human can learn from reading books without infringing copyright, why can't a machine similarly learn from training data? However, when that training data is comprised of data downloaded from the internet, copies are necessarily created in the process of training a machine learning model.

It is this intermediate copying that differentiates machine learning from human learning and explains why the former implicates copyright law. The Ninth Circuit in *Sega Enterprises v. Accolade, Inc.* held that the intermediate cop-

[49] 991 F.2d 511, 519 (9th Cir. 1993). *But see* Cartoon Network LP, LLLP v. CSC Holdings, Inc., 536 F.3d 121, 130 (2d Cir. 2008) (holding that a period of 1.2 seconds was not sufficiently fixed to constitute copyright infringement).

[50] *See Stern Elecs.*, 669 F.2d at 856; *MAI Sys. Corp.*, 991 F.2d at 519; John Cruickshank, *The Difference Between RAM and Hard Drive*, TECHTORIUM (May 8, 2020, 8:49 PM), https://techtorium.ac.nz/the-difference-between-ram-and-hard-drive.

[51] *See, e.g.*, *Working with Cloud Storage*, GOOGLE CLOUD (Nov. 16, 2020), https://cloud.google.com/ai-platform/training/docs/working-with-cloud-storage.

[52] *See* IBM Cloud Education, *Cloud Storage*, IBM (June 24, 2019), https://www.ibm.com/cloud/learn/cloud-storage.

[53] 17 U.S.C. § 101; *see SSD vs. HDD: Which is Best for You?*, INTEL, https://www.intel.com/content/www/us/en/products/docs/memory-storage/solid-state-drives/ssd-vs-hdd.html.

ying of protected computer code could constitute copyright infringement, regardless of whether the end product of the copying also infringed.[54] Applying that precedent, the District of Nevada in *Tiffany Design, Inc. v. Reno-Tahoe Specialty* found that intermediate copying through the scanning of protected photographs constituted copyright infringement as a matter of law, without even determining whether the end product, an artistic depiction of the Las Vegas Strip, was substantially similar.[55]

Despite the intermediate copying, the defendants in *Sega* and *Sony Computer Entertainment v. Connectix Corp.*, a more recent computer case also involving the intermediate copying of protected code, prevailed under the fair use defense.[56] The fair use doctrine, although riddled with imperfections, currently represents an AI developer's best chance at defeating a claim of copyright infringement in court.

C. THE FAIR USE DOCTRINE

Courts have applied the fair use doctrine to find non-infringement in technology cases. Codified in 17 U.S.C. § 107, the fair use doctrine is an affirmative defense to a claim of copyright infringement.[57] The analysis considers four factors:

> (1) the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes;
>
> (2) the nature of the copyrighted work;
>
> (3) the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and
>
> (4) the effect of the use upon the potential market for or value of the copyrighted work.[58]

The fair use doctrine has led to unexpected results when these four factors are applied to new technological contexts.[59]

For example, the Ninth Circuit in *Sega* ruled that Accolade's reverse engineering of Sega's Genesis video game programs qualified as fair use even

[54] 977 F.2d 1510, 1519 (9th Cir. 1992).

[55] 55 F. Supp. 2d 1113, 1121 (D. Nev. 1999).

[56] 17 U.S.C. § 107; *see Sega*, 977 F.2d at 1514; Sony Computer Ent., Inc. v. Connectix Corp., 203 F.3d 596, 598 (9th Cir. 2000).

[57] 17 U.S.C. § 107.

[58] *Id.*

[59] *See Sega*, 977 F.2d at 1527 ("We are not unaware of the fact that to those used to considering copyright issues in more traditional contexts, our result may seem incongruous at first blush.").

though the intermediate copying of protected code was necessary.[60] The court stated that the first factor, "the purpose and character of the use," weighed in favor of Accolade because Accolade's ultimate goal was the release of Genesis-compatible games for sale—a "legitimate, essentially non-exploitative purpose."[61] Accolade's copying of protected code was necessary for extracting the functional requirements for Genesis compatibility so that it could make its existing games compatible with the Genesis console.[62]

Relatedly, under the second factor, "the nature of the copyrighted work," Sega's video game programs were afforded a lower degree of protection than traditional literary works because they contained unprotected, functional aspects like Genesis compatibility.[63] The second factor analysis essentially incorporated the idea-expression dichotomy of copyright law—that ideas and functional concepts are not protected by copyright.[64] Additionally, the fourth factor, the effect of the use on the market, also weighed in favor of Accolade because its entry into the market for Genesis-compatible games would promote creative expression in video games.[65] Finally, under the third factor, the amount and substantiality of the copied portion, copying of the entire code was not enough to outweigh the effect of the other three factors.[66] Thus, the Ninth Circuit in *Sega* ruled that Accolade's reverse engineering was fair use.

More recently, the first factor of the fair use analysis has tended to overshadow the second and third factors because of the introduction of the doctrine of transformative use.[67] As explained by Judge Leval in his seminal law review article, central to the first factor of the fair use analysis is whether the challenged use is *transformative*, that is, whether the original is used as "raw material, transformed in the creation of new information, new aesthetics, new insights and understandings."[68] Since its introduction, the percentage of transformative use decisions in fair use cases has increased from 8% in 1991 to

[60] *Id.* at 1520.

[61] *Id.* at 1522–23.

[62] *Id.*

[63] *Id.* at 1524–26.

[64] *See id.* at 1519–20; 17 U.S.C. § 102(b).

[65] *Sega*, 977 F.2d at 1524.

[66] *Id.* at 1526–27.

[67] Jane C. Ginsburg, *Fair Use in the United States: Transformed, Deformed, Reformed?*, 2020 SING. J. LEGAL STUD. 265, 277 n.56 (2020).

[68] Pierre N. Leval, Commentary, *Toward a Fair Use Standard*, 103 HARV. L. REV. 1105, 1111 (1990).

around 90% in recent years.[69] And out of a large sample of decisions, 94% that found transformative use also ultimately found fair use.[70]

Relevant to this Note's data mining context, this transformative analysis has also been applied to cases involving internet search engines that functionally utilize text and images. For instance, the Ninth Circuit in *Perfect 10, Inc. v. Amazon.com, Inc.* held that the use of copyrighted images to display search engine results was fair use because the use of thumbnails was transformative.[71] While the images originally served "entertainment, aesthetic, or informative function[s]," the search engine transformed the image into a "pointer directing a user to a source of information."[72] Similarly, the Second Circuit in *Authors Guild, Inc. v. Google, Inc.* held that Google's digitizing of copyrighted texts was fair use because the Google Books search engine was highly transformative.[73] Although Google scanned full copyrighted texts, it displayed only snippets of text, which acted as "pointers directing users to a broad selection of books."[74] These cases suggest that courts are likely to find fair use exceptions to copyright infringement where the end product serves a functional purpose and provides social utility.

For an AI developer, the search engine cases provide some precedent for structuring a fair use defense. If internet search engines were determined to be transformative, it is easy to imagine how AI products could qualify as highly transformative uses of expressive data (i.e., surgery-assistant robots, self-driving cars, translation services). Images, video, and text are fed as "raw material" to train models that teach machines to make decisions as humans do—creating "new insights and understandings."[75] Even the massive scale reproductions of images or text performed on the computer are not prohibitive as Google scanned "tens of millions of books" for use in its search engine.[76] And *Authors Guild, Inc. v. Google, Inc.* showed that a defendant's highly commercial nature does not outweigh a product's transformative value.[77] Thus, if an AI developer were downloading thousands of copyrighted images to train a computer vision

[69] Jiarui Liu, *An Empirical Study of Transformative Use in Copyright Law*, 22 STAN. TECH. L. REV. 163, 174 (2019).

[70] *Id.* at 180.

[71] 508 F.3d 1146, 1167 (9th Cir. 2007).

[72] *Id.* at 1165.

[73] 804 F.3d 202, 216 (2d Cir. 2015).

[74] Authors Guild, Inc. v. Google Inc., 954 F. Supp. 2d 282, 291 (S.D.N.Y. 2013), *aff'd*, 804 F.3d at 202.

[75] Leval, *supra* note 68, at 1111.

[76] *Authors Guild*, 804 F.3d at 206.

[77] *Id.* at 209.

model, perhaps she too could rely on the fair use defense, even if she intended to make a commercial product out of it.

However, the fair use doctrine fluctuates in its friendliness to technology.[78] In a more recent search engine case, the Ninth Circuit held that Zillow's use of copyrighted photographs in its apartment listing searches was not fair use.[79] Zillow's database featured photos of artfully-designed rooms that could be sorted using "various criteria, like room type, style, cost, and color."[80] Although the search engine made the photographs functionally searchable, the court found that it did not fundamentally change their original purpose to "artfully depict rooms and properties," and thus preserved the original photos' "inherent character."[81] Additionally, because the plaintiff was "actively exploring" a market for licensing its photographs, the fourth factor weighed against Zillow.[82] Similarly, the Second Circuit in *Fox News Network, LLC v. TVEyes, Inc.* found no fair use for a TV clip search engine that was "somewhat transformative" but usurped a potential market for Fox to license its works.[83]

These recent cases suggest that the fair use defense may not be so straightforward when applied to AI technologies.[84] Although it may have once been enough to lean on the first factor and point out that an end use is transformative, *VHT, Inc.* and *TVEyes* show that courts may heavily consider the potential effect that the use has on licensing markets.[85] Thus, a developer may be penalized under the fair use analysis for depriving a copyright owner of the opportunity to license her works to be used in training datasets. And the case law illustrates that copyrighted photographs do not receive weaker protection

[78] *See* Michael W. Carroll, *Copyright and the Progress of Science: Why Text and Data Mining is Lawful*, 53 U.C. DAVIS L. REV. 893, 936 (2019).

[79] VHT, Inc. v. Zillow Grp., Inc., 918 F.3d 723, 744 (9th Cir 2019). Another recent case weighs unfavorably on the use of copyrighted photos for informational purposes. In *Brammer v. Violent Hues Productions*, the Fourth Circuit held that a film website's use of a copyrighted photograph, found from a Google Images search, was not fair use. 922 F.3d 255, 269 (4th Cir. 2019). The defendant's claim that the photograph was used to provide information regarding a tourist attraction was not convincing, otherwise "virtually all illustrative uses of photography would qualify as transformative." *Id.* at 264. The court also noted that photographs have long received "thick copyright protection" even though they "capture images of reality." *Id.* at 266–67.

[80] *VHT, Inc.*, 918 F.3d at 730.

[81] *Id.* at 742.

[82] *Id.* at 744.

[83] *See* 883 F.3d 169, 178–80 (2d Cir. 2018).

[84] *See VHT, Inc.*, 918 F.3d at 740 ("While we can discern certain animating principles bridging cases in this area, the doctrine has hardly followed a straight, or even slightly curved, line.").

[85] *See id.* at 744; *TVEyes*, 883 F.3d at 180–81.

even if used for "informational" purposes.[86] Even if an AI developer tries to invoke the idea-expression dichotomy and argues that she uses the informational aspects of copyrighted photographs to teach machine learning models how to identify objects, a court may find that she did not alter the work's original purpose of depicting such objects in the process of compiling her training data set.

Rather than relying on a fair use defense, it would be more ideal for enterprising AI developers if a clear safe harbor in copyright law allowed the open use of expressive works in training data.

III. AN ARGUMENT FOR A DATA MINING SAFE HARBOR

Fundamentally, data mining is not copyright infringement but rather involves the lawful copying of unprotected material forms. Practically, providing a safe harbor for data mining on top of the existing fair use doctrine would provide legal certainty to innovators who may currently be deterred by the threat of litigation and hefty statutory fines.

A. DATA MINING IS NOT COPYRIGHT INFRINGEMENT

Not all copying is copyright infringement.[87] Professor Abraham Drassinower uses *Baker v. Selden* to illustrate the distinction between a copyrighted work and its material form, arguing that not all uses of the work's material form are acts of infringement.[88] In the classic Supreme Court case, Selden obtained a copyright on his book explaining the operation of a novel accounting system, which included accounting forms as part of the explanation.[89] Baker then used a similar system, which included copying the forms, as explained and illustrated in Selden's books.[90] The forms were unique in that they both described the accounting system and were also used to perform the accounting.[91] While copyright law protects the use of the forms as an explanation, it does not protect the use of the forms as an invention (which falls within the realm of patents).[92] The defendant escaped copyright liability because he used the plaintiff's forms as aspects of a novel accounting system, but not as aspects of an explanation of the accounting system.[93]

[86] *See* Brammer v. Violent Hues Prods., LLC, 922 F.3d 255, 267–69 (4th Cir. 2019).

[87] Feist Publ'ns, Inc. v. Rural Tel. Serv. Co., 499 U.S. 340, 361 (1991).

[88] *See* DRASSINOWER, *supra* note 10, at 88–100.

[89] Baker v. Selden, 101 U.S. 99, 100 (1879).

[90] *Id.* at 101.

[91] DRASSINOWER, *supra* note 10, at 89.

[92] *Id.*

[93] *Id.* at 93.

Thus, copyright infringement requires not just copying of a work's material form, but also the unauthorized use of the work for its expressive purpose.[94] By the same logic, merely technical or non-communicative uses are not uses of a work for its expressive purpose and therefore are not copyright infringement.[95] Because they are not expressive uses of a work at all, technical and noncommunicative uses need not even be subject to a fair use analysis.[96]

Applying Drassinower's logic to data mining, this Note argues that downloading images from the internet for use in training data is not copyright infringement, but rather lawful copying of the works' material forms. An AI developer does not redistribute, or re-communicate, the copyrighted images to the public but instead uses them to train a machine learning model. Because copyright protection does not extend to the material forms of works themselves, the simple act of downloading images does not encroach upon a protected use of copyrighted works. By reading too closely into fixation requirements for intermediate copying in software cases and too eagerly applying the fair use analysis, courts have overlooked the simple explanation that some acts of reproduction are not copyright infringement because they are non-expressive reproductions of material forms only, analogous to Baker's copying of Selden's accounting forms.[97]

Thus, imposing copyright liability on developers who use expressive works to train their functional models would be an overreach of creators' rights. Copyright law is intended to protect creative and expressive works.[98] The idea-expression dichotomy limits copyright protection to the expressive elements of a work and not the functional ideas contained within.[99] In building a machine learning training set, the developer is not interested in reproducing the expressive works but is instead interested in the functional content contained within the material forms. For example, NLP developers may be interested in using literary works as training data to extract foundational patterns of human

[94] *See id.* at 94.

[95] *Id.* at 87–88, 100–102.

[96] *Id.* at 101.

[97] *See* Baker v. Selden, 101 U.S. 99, 107 (1879).

[98] *See* 17 U.S.C. § 102.

[99] *See* Mazer v. Stein, 347 U.S. 201, 217 (1954) ("[A] copyright gives no exclusive right to the art disclosed; protection is given only to the expression of the idea—not the idea itself."); Harper & Row Publishers, Inc. v. Nation Enters., 471 U.S. 539, 556 (1985) ("[C]opyright's idea/expression dichotomy 'strike[s] a definitional balance between the First Amendment and the Copyright Act by permitting free communication of facts while still protecting an author's expression.'" (quoting Harper & Row Publishers, Inc. v. Nation Enters., 723 F.2d 195, 203 (2d Cir. 1983), *rev'd*, 471 U.S. at 539)).

speech.[100] And in computer vision, developers use videos and photographs of busy city streets to teach machines how to identify pedestrians.[101] In neither case is the machine copying the expressive character of the works (i.e., the expression in the writing itself or the artful depiction of city streets in a photograph).

Additionally, allowing the open use of copyrighted material as training data would further the broader goals of intellectual property law. As Judge Leval noted in *Authors Guild*, "[t]he ultimate goal of copyright is to expand public knowledge and understanding . . . . [T]he ultimate, primary intended beneficiary [of copyright] is the public."[102] A safe harbor in copyright law for data mining would provide public benefit by stimulating innovation. New applications of artificial intelligence are constantly conjured up: computer vision technology can help drones spot wildfire hazards in California,[103] detect bias and trends on TV,[104] and catch illegal elephant poachers in Africa.[105] AI-driven tools like Grammarly and Adobe Sensei even assist writers and artists in the creation of new content.[106] These beneficial applications could be more numerous and of higher quality if potential innovators had open access to copyrighted works as training data.[107] Allowing copyright to hinder the growth of beneficial technology that does not infringe upon creators' rights would ultimately impede the goals of intellectual property law.

[100] *See* Ana Cristina Mendes & Cláudia Antunes, *Pattern Mining with Natural Language Processing: An Exploratory Approach*, *in* MACHINE LEARNING AND DATA MINING IN PATTERN RECOGNITION 266, 266 (Petra Perner ed., 2009) ("Articles from a Portuguese newspaper are the input . . . . Results . . . provided several evidences about the structure of the language.").

[101] *See* Antonio Brunetti, Domenico Buongiorno, Gianpaolo Francesco Trotta & Vitoantonio Bevilacqua, *Computer Vision and Deep Learning Techniques for Pedestrian Detection and Tracking: A Survey*, 300 NEUROCOMPUTING 17, 20–21 (2018).

[102] 804 F.3d 202, 212 (2d Cir. 2015).

[103] John McCormick, *California Utilities Hope Drones, AI Will Lower Risk of Future Wildfires*, WALL ST. J. (Sept. 11, 2020, 5:30 AM), https://www.wsj.com/articles/california-utilities-hope-drones-ai-will-lower-risk-of-future-wildfires-11599816601?mod=hp_minor_pos4.

[104] Sherin Shibu, *Who are the Biggest Attention Hogs on Cable News? Ask this new AI Tool*, PCMAG (Sept. 1, 2020), https://www.pcmag.com/news/who-are-the-biggest-attention-hogs-on-cable-news-ask-this-new-ai-tool.

[105] Simorin Pinto, *AI-powered Camera to Stop Illegal Poaching*, INNOVATION ENTER. CHANNELS, https://channels.theinnovationenterprise.com/articles/ai-powered-camera-to-stop-illegal-poaching [https://perma.cc/93NU-43LJ].

[106] *See, e.g.*, Grammarly, *How We Use AI to Enhance Your Writing | Grammarly Spotlight*, GRAMMARLY BLOG (May 17, 2019), https://www.grammarly.com/blog/how-grammarly-uses-ai; Marc DeAngelis, *Adobe Premiere Pro Can Automatically Reframe Your Videos*, ENGADGET (Sept. 13, 2019), https://www.engadget.com/2019-09-13-adobe-premiere-pro-reframes-video.html.

[107] *See infra* Section III.B.3.

B. A SAFE HARBOR TO FIX LEGAL UNCERTAINTY

Because data mining falls outside the scope of copyright, a safe harbor for such uses should be enacted. Clarifying the law would promote innovation by providing legal certainty for smaller innovators.

1. *The fair use doctrine should not be applied to data mining because it would result in further stretching of the unpredictable doctrine.*

Most copyright stakeholders argue that fair use is currently sufficient because the "flexible doctrine . . . is capable of adapting to the use of copyrighted works in an AI context."[108] In practice, though, fair use has been referred to as "the most troublesome [doctrine] in the whole law of copyright" and described as a "billowing white goo."[109] Because data mining is not copyright infringement, the inquiry should end before the courthouse doors, rendering the nebulous fair use analysis moot. A safe harbor would not displace the fair use defense in the United States for all AI application but would provide an additional layer of certainty on top of the doctrine of functional uses. Certainty in the law is desirable because even the possibility of litigation may be a deterrent for smaller creators.[110]

Conceptually, the fair use doctrine should not even apply to data mining because downloading data from the internet for use in training data is not a use of protected works.[111] Drassinower differentiates this nonuse from fair use based on whether the use is communicative.[112] The fair use doctrine protects communicative uses of copyrighted works and appeals to the "equality as authors"—the defendant reproduces the plaintiff's work as a "reasonably necessary aspect of the defendant's own authorial engagement."[113] Fair use is rightly invoked for communicative uses like parody, art, and criticism.[114] Conversely,

[108] *See* U.S. PAT. & TRADEMARK OFF., *supra* note 6, at 26; Stan Adams, *Comments On the USPTO's Intellectual Property Protection for Artificial Intelligence Innovation*, CTR. FOR DEMOCRACY & TECH. (Jan. 16, 2020), https://cdt.org/insights/comments-on-the-usptos-intellectual-property-protection-for-artificial-intelligence-innovation ("[F]air use offers advantages over other possible legal mechanisms for allowing the use of copyrighted works in the context of machine learning . . . . [T]he resulting [EU] exception for text and data mining (TDM) is so rigid and restrictive as to prevent many beneficial uses of datasets.").

[109] VHT, Inc. v. Zillow Grp., Inc., 918 F.3d 723, 739 (9th Cir. 2019) (quoting Monge v. Maya Mags., Inc., 688 F.3d 1164, 1170–71 (9th Cir. 2012)).

[110] *See infra* Section III.B.2.

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

[112] *See* DRASSINOWER, *supra* note 10, at 100–01.

[113] *Id.* at 108.

[114] *See, e.g.*, Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 572 (1994); Blanch v. Koons, 467 F.3d 244, 246 (2d Cir. 2006); Salinger v. Random House, Inc., 811 F.2d 90, 92 (2d Cir. 1987).

nonuse stems out of the distinction between a work and its material form; the defendant is not liable because she only makes use of the material form of the work and not the work as a communicative act.[115] Nonuses include search engine thumbnail reproductions, temporary copies made during internet browsing, and digital copies of student papers made to detect plagiarism.[116] Courts confusing nonuses for fair uses have contributed to the confusion in the fair use doctrine.

Application of the fair use defense to digital technologies has also stretched the doctrine to the detriment of creators in general. As Professor Jiarui Liu has observed, lower courts stretch transformative use to new fact patterns, contributing to a "slippery slope progression."[117] He argues that the trend in intermediate copying decisions perfectly illustrates the phenomenon: logic applied to copies made and deleted during the reverse engineering of video games was stretched to justify search engines that continuously store and display verbatim copies.[118] Instead of relying on an overstretching of the fair use doctrine, courts should have permitted search engine uses of copyrighted works as nonuses of the protected works.[119]

Proponents of justifying data mining with fair use must also keep in mind that the fair use doctrine applies across the board to decisions involving literature, art, and music. Fair use decisions have wide-reaching implications for other creative mediums. For example, the Supreme Court suggested that a musical parody was unlikely to act as a substitute for an original song because the two works serve different market functions.[120] Expanding on that logic, the Second Circuit in *Cariou v. Prince* pointed to the differences in wealth in consumers to justify a finding of fair use.[121] In *Cariou*, the two artists served two different markets of art collectors: while the plaintiff earned just thousands of dollars in royalties and sold only to personal acquaintances, the defendant sold millions of dollars' worth of art to celebrities.[122] Ultimately, the reasoning used to justify digital technologies under the fair use doctrine could narrow the rights of authors in other areas.

[115] DRASSINOWER, *supra* note 10, at 108–09.

[116] *See id.* at 87, 102–04.

[117] Liu, *supra* note 69, at 211.

[118] *Id.*

[119] DRASSINOWER, *supra* note 10, at 100–03.

[120] Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 572 (1994).

[121] *See* Cariou v. Prince, 714 F.3d 694, 709 (2d Cir. 2013); Liu, *supra* note 69, at 199–200.

[122] *Cariou*, 714 F.3d at 709.

Liu also argues that the transformative analysis under the fair use doctrine has become a way for courts to decide new policy concerns under the "impression of *stare decisis*."[123] Given the malleability of the doctrine and the wide scope of case law, courts usually have little difficulty in finding a precedent that justifies fair use decisions.[124] This constant expansion of transformative use creates further uncertainty in copyright law. Rather than rely on district courts to dictate important policy issues involving emerging technologies, like the training of machine learning models with copyrighted data, Congress should be more proactive in enacting safe harbors for new technologies.

Practically, the fair use doctrine is an affirmative defense that must be raised during litigation.[125] And, because fair use is an affirmative defense, a claim of copyright infringement through data mining could survive a motion to dismiss because intermediate copying is infringement as a matter of law.[126] The fair use case law provides conflicting guidance for innovators to try to understand how their technology will be viewed by the courts.[127] The unpredictability of the fair use doctrine is especially apparent with technology cases; the acclaimed copyright scholar William F. Patry once predicted that Google Books was unlikely to stand the fair use test in a now-deleted blog post (he has since joined Google's legal team).[128] If great legal minds cannot make sense of the fair use doctrine, how can innovators be expected to rely on the defense when developing new technologies? There are already examples of this uncertainty among smaller innovators as discussed in the next Section.[129]

2. *Uncertainty in copyright law disproportionately handicaps new innovators.*

A safe harbor in copyright law for data mining is further warranted because current uncertainty in the law disproportionately handicaps small innovators. At the "Copyright in the Age of Artificial Intelligence" conference hosted by

[123] Liu, *supra* note 69, at 172.

[124] *Id.*

[125] *See Campbell*, 510 U.S. at 599.

[126] *See* Tiffany Design, Inc. v. Reno-Tahoe Specialty, Inc., 55 F. Supp. 2d 1113, 1121 (D. Nev. 1999) (holding that intermediate copying is copyright infringement as a matter of law).

[127] *See supra* Section II.C.

[128] Liu, *supra* note 69, at 169.

[129] *See, e.g.*, Peter Ned, Comment to *In the US, is it Illegal to Train Neural Networks Using Copyrighted Images?*, QUORA (July 1, 2017), https://www.quora.com/In-the-US-is-it-illegal-to-train-neural-networks-using-copyrighted-images; farsass, Comment to *Is it Legal to Use Copyright Material as Training Data?*, REDDIT (July 1, 2016, 8:33 AM), https://www.reddit.com/r/MachineLearning/comments/4qrgh8/is_it_legal_to_use_copyright_material_as_training; bluboxsw, Comment to *[D] Are There Any Legal Issues with Training Machine Learning Models on Copyrighted Content?*, REDDIT (July 12, 2019, 10:50 AM), https://www.reddit.com/r/MachineLearning/comments/cc76us/d_are_there_any_legal_issues_with_training.

the U.S. Copyright Office and the World Intellectual Property Organization in 2020, Vanessa Bailey, the global director of IP policy for Intel Corporation, stated, "[C]opyright laws are still adequate . . . . [W]e're doing fine with what we have and . . . smart lawyers are figuring things out."[130] This statement illustrates the wealth of resources that large tech companies enjoy: access to enormous datasets and expensive legal teams. The argument would be different from the perspective of a smaller innovator because copyright law can create disparities in access to data, ability to litigate, and deterrent effects against using larger datasets.

Copyright law contributes to the disproportionate access to valuable training data between dominant tech companies and smaller innovators. Large tech platforms like YouTube and Meta operate with terms of service that provide them with access to copyright-protected text and data uploaded to their servers.[131] This data presents a treasure trove for training machine learning models. For example, Meta uses its data from over 2 billion users to calibrate newsfeeds, generate text for the visually impaired, and power its facial recognition technology.[132] Even when large companies do not have built-in systems for acquiring data, they can buy large datasets.[133] These acquisitions of large datasets can be extremely expensive; one such partnership IBM undertook for access to oncology data cost $50 million.[134] And though some large tech companies have released open source algorithms, it is rarer for them to release the underlying datasets.[135]

[130] U.S. COPYRIGHT OFF., COPYRIGHT IN THE AGE OF ARTIFICIAL INTELLIGENCE 3, 294 (2020) (transcript of symposium).

[131] *See Terms of Service*, YOUTUBE (Nov. 18, 2020), https://www.youtube.com/static?template=terms [https://perma.cc/3NTH-KB9V] ("By providing Content to the Service, you grant to YouTube a worldwide, non-exclusive, royalty-free, sublicensable and transferable license to use that Content (including to reproduce, distribute, prepare derivative works, display and perform it) . . . ."); *Terms of Service*, FACEBOOK (Oct. 22, 2020), https://www.facebook.com/terms.php [https://perma.cc/Z7MS-4D3R]("Specifically, when you share, post, or upload content that is covered by intellectual property rights on or in connection with our Products, you grant us a non-exclusive, transferable, sub-licensable, royalty-free, and worldwide license to host, use, distribute, modify, run, copy, publicly perform or display, translate, and create derivative works of your content . . . .").

[132] Amanda Levendowski, *How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem*, 93 WASH. L. REV. 579, 606–07 (2018).

[133] *Id.* at 606.

[134] *Id.* at 608.

[135] *Id.* at 599.

Thus, smaller innovators do not have equivalent access to datasets, which are an important resource for developing good algorithms.[136] Although open source datasets do exist, they may not be comparable in size or may be prone to bias.[137] However, if a safe harbor in copyright law clarified that mining copyrighted content for machine learning training purposes was legal, these smaller innovators would enjoy freer access to valuable data, allowing them to develop innovative products that could allow them to better compete against tech giants. Increased competition would benefit the public by providing access to more numerous and higher quality AI products.

Notwithstanding the data that large tech companies legally have access to, it is "plausible, if not probable, that dominant AI players create unauthorized copies of protectable works to use as training data for AI systems."[138] Even if large tech companies run afoul of copyright law in training their machine learning algorithms, they are better equipped with lawyers to defend themselves from liability. As illustrated previously, the doctrine of fair use is highly fact-specific and arguably unpredictable, which creates room for smart (i.e., creative, expensive) lawyering.[139] The tech giant Google secured a victory in the landmark fair use case *Authors Guild v. Google, Inc.* after ten years of litigation.[140] Conversely, litigation costs can drain the bank accounts of startups and lead them to cease operations.[141] The support of affluent institutions in copyright cases like *Cariou v. Prince* further hints at the influence of wealth in securing favorable verdicts.[142]

[136] *See* Steven Levy, *Inside Facebook's AI Machine*, WIRED (Feb. 23, 2017, 12:00 AM), https://www.wired.com/2017/02/inside-facebooks-ai-machine/ ("One huge factor in building machine learning systems is getting quality data—the more the better.").

[137] *See id.* ("When you have over a billion people interacting with your product every day, you collect a *lot* of data for your training sets . . . ."); *infra* Section III.B.3.

[138] Levendowski, *supra* note 132, at 599.

[139] *See supra* Part II, Section III.B.1.

[140] *See* Adam Liptak & Alexandra Alter, *Challenge to Google Books is Declined by Supreme Court*, N.Y. TIMES (Apr. 18, 2016), https://www.nytimes.com/2016/04/19/technology/google-books-case.html.

[141] *See* Moran, *supra* note 7 ("Litigation is expensive—no matter how speculative the claims against you nor how worthy your position . . . . With our company ensnared by this legal battle, we have been unable to raise another round of funding to fuel our development and marketing efforts. Our bank account is running out, and we must cease operations in the new year.").

[142] *See* Douglas W. Kenyon & Stephen P. Demm, *Supreme Court Denies Cert in Cariou Fair Use Case: What Next?*, INTELL. PROP. MAG., Feb. 2014, at 71, 72 ("And in an unusual step at the district court level, the Andy Warhol Foundation and 30 other arts groups filed an *amicus brief* . . . ."); Hrag Vartanian, *Cariou v. Prince Isn't Over: Orgs Rep'ing 45,000 Creatives File Brief in Support of Cariou*, HYPERALLERGIC (Dec. 20, 2013), https://hyperallergic.com/99668/cariou-v-prince-isnt-over-orgs-reping-45000-creatives-file-brief-in-support-of-cariou ("Amici submit this brief primarily in opposition to the amicus brief filed by the Andy Warhol Foundation

Coupled with the expense of litigation itself, the U.S. remedial regime of copyright law also presents a significant deterrent for new innovators. Statutory damages range from $200 to $150,000 per work, depending on the willfulness of infringement.[143] Training a model from scratch can require millions of data points, and even fine-tuning a machine learning model can require thousands of data points.[144] Because of the large quantity of data needed to develop a machine learning model, a developer could potentially be on the hook for massive fines. This U.S. remedial regime of statutory damages creates the potential for crushing liability, especially when amplified by the massive amounts of data required to train a machine learning model.[145] The legal uncertainty and mere threat of exorbitant fines surrounding the use of copyrighted data would be enough to deter smaller actors from creating and using potentially valuable datasets.

The effect of this legal uncertainty is readily observed in practice. A look at user forums such as Quora and Reddit illustrates caution against using copyrighted data for training machine learning models:

> "Sorry, I can't answer the question. And I doubt anyone can. It probably requires a series of multi-million lawsuits involving herds of lawyers to find out. To be on the safe side don't use copyrighted images, at least for now."—Peter Ned[146]
>
> "You can find a lawyer to argue for or against this being a case of copyright violation depending on how much you pay him."—farsass[147]
>
> "It is still derivative work, which is covered under law. When the $$$ you make exceeds the $$$ to sue you, expect to get sued."—bluboxsw[148]

and other elite foundations and museums who do not represent the views of working artists." (emphasis omitted)).

[143] 17 U.S.C. § 504.

[144] Brown et al., *supra* note 9.

[145] *See* Menell, *supra* note 40, at 268.

[146] Peter Ned, Comment to *In the US, is it Illegal to Train Neural Networks Using Copyrighted Images?*, QUORA (July 1, 2017), https://www.quora.com/In-the-US-is-it-illegal-to-train-neural-networks-using-copyrighted-images.

[147] farsass, Comment to *Is it Legal to Use Copyright Material as Training Data?*, REDDIT (July 1, 2016, 8:33 AM), https://www.reddit.com/r/MachineLearning/comments/4qrgh8/is_it_legal_to_use_copyright_material_as_training.

[148] bluboxsw, Comment to *[D] Are There Any Legal Issues with Training Machine Learning Models on Copyrighted Content?*, REDDIT (July 12, 2019, 10:50 AM), https://www.reddit.com/r/MachineLearning/comments/cc76us/d_are_there_any_legal_issues_with_training.

A developer seeking to gather a wide expanse of training data could be deterred by such general uncertainty.

Relatedly, questions of access influence how scholars conduct their research. NLP techniques can be used to study literary works.[149] However, researchers may limit research questions based on the availability of texts. Ethan Reed, a digital humanities researcher interested in articulations associated with systemic injustice, laments that copyright plays "an enormous role in determining the initial paths in my scholarly decision-making process."[150] In one example, he limited the scope of an NLP research project to just three books of poetry from 1969 because of copyright concerns.[151] In explaining this decision, he highlights an additional problem of reproducibility of research.[152] Even if scholars can share results of NLP analyses through transformative, non-consumptive use, they cannot share the copyrighted corpora from which the insights came from.[153] Copyright limits the scope of computational humanities research and potentially stymies socially valuable insights that can be derived from contemporary works.

AI practitioner Arjan Wijnveen also describes another copyright problem that developers face—the decay in public datasets used to train models.[154] Public datasets like ImageNet provide huge volumes of annotated data for developers to train on but are based on public image or video hosting sites.[155] ImageNet does not make images publicly available in their original resolutions because they might be subject to copyright, instead providing thumbnails and URLs.[156] When an image or video that is part of an annotated dataset is taken down, developers are "out of luck."[157] An obvious solution would be to store

[149] *See, e.g.*, Senja Pollak, Matej Martinc & Katja Mihurko Poniz, *Natural Language Processing for Literary Text Analysis: Word-Embeddings-Based Analysis of Zofka Kveder's Work*, 2607 CEUR WORKSHOP PROC. 33 (2020) (using NLP to gain insights in feminine literary history).

[150] Ethan Reed, *First Steps with NLP and a Collection of Amiri Baraka's Poetry*, SCHOLAR'S LAB (Nov. 30, 2017), https://scholarslab.lib.virginia.edu/blog/first-steps-with-nlp-and-a-collection-of-amiri-barakas-poetry ("Though conceptually unglamorous, basic questions of access have played an enormous role in determining the initial paths in my scholarly decision-making process.").

[151] *Id.*

[152] *Id.*

[153] *Id.*

[154] Arjan Wijnveen, *How Copyright is Causing a Decay in Public Imagesets*, MEDIUM (Nov. 28, 2016), https://medium.com/@arjanwijnveen/how-copyright-is-causing-a-decay-in-public-datasets-f760c5510418.

[155] *Id.*

[156] *Download FAQ*, IMAGENET, http://image-net.org/download-faq ("The images in their original resolutions may be subject to copyright, so we do not make them publicly available on our server.").

[157] Wijnveen, *supra* note 154.

a cached copy of the images, but Wijnveen believes that this fix is prohibited by copyright law.[158] Thus, vague copyright laws can undermine the usefulness of public datasets.

3. Licensing is an overreach of authors' rights and could propagate bias.

Some commentators argue in favor of licenses for AI training data used in commercial applications.[159] However, requiring the licensing of training data would be burdensome to developers and an overreach of authors' rights. Data for machine learning training sets is usually scraped from the internet, and this process is automated because of the large quantities of data needed.[160] Given the massive size of datasets and the automation of the scraping process, it would be extremely burdensome for developers to go through scraped data, determine what is copyrighted, and request permission from each creator.

Most importantly, because reproductions in data mining are fundamentally not infringement, requiring compensation for these merely technical and non-communicative uses would be an overreach of authors' rights. It does not matter if a market for training data exists or could exist.[161] In *Baker v. Selden*, the fact that a market for accounting forms existed did not change the finding that Baker's use of Selden's accounting forms was an unactionable use under copyright law.[162] Allowing an author to charge for all reproductions of the material form of her work would overextend her claim to rights over her work as a material thing, while copyright applies only to works as communicative acts.[163]

Some may also argue that developers have already adjusted to copyright uncertainty by using licensed or safe datasets. For example, developers can circumvent liability by training on datasets made available under creative commons (CC) licenses[164] or by training on corpora of text from websites, like Wikipedia.org, that allow free access, copying, and distribution.[165] However,

[158] *Id.*

[159] *See* U.S. Pat. & Trademark Off., *Artificial Intelligence: The Ins and Outs of Copyright and AI*, VBRICK REV, at 43:00 (Jan. 31, 2019), https://rev-vbrick.uspto.gov/#/videos/d6e591c3-64cf-4d74-ab35-9f387a2da4b2 (highlighting the views of Mary Ransenberger, Executive Director of the Authors Guild, who proposes solution of a collective licensing system).

[160] *See* Casper Hansen, *Web Scraping For Machine Learning - With SQL Database*, MACH. LEARNING FROM SCRATCH (Dec. 4, 2019), https://mlfromscratch.com/web-scraping-machine-learning/#.

[161] *See* DRASSINOWER, *supra* note 10, at 102.

[162] *Id.*

[163] *Id.* at 99.

[164] *See, e.g.*, *Updated Dataset*, YOUTUBE-8M, https://research.google.com/youtube8m/download.html ("The dataset is made available by Google LLC. under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.").

[165] *See, e.g.*, *Wikipedia:Copyrights*, WIKIPEDIA, https://en.wikipedia.org/wiki/Wikipedia:Copyrights ("Wikipedia content can be copied, modified, and redistributed *if and only if* the

Professor Amanda Levendowski points out an important problem that arises from the use of safe datasets—the propagation of AI bias.[166] She argues that copyright law encourages AI creators to use "easily available, legally low-risk works," such as public domain and creative commons-licensed works, for training data.[167] The use of this safe data is problematic because the sources are often biased.[168] For example, most public domain literary works that could be used for NLP applications "were published prior to 1923, back when the 'literary canon' was wealthier, whiter, and more Western than it is today," and thus an AI model trained on these works would reflect the biases of the time.[169] Additionally, training data from CC-licensed websites like Wikipedia can reflect the biases inherent in editors; for example, only 10–20% of Wikipedia editors are women.[170] This gender disparity contributes to male-centric representations of facts, which can be propagated in biased models.[171]

In general, larger and more diverse datasets create better models.[172] And, by clearly broadening the scope of data that can be mined to include copyrighted works, reliance on safe, but biased, data is weakened. Models trained on diverse, but potentially copyrighted, data will be more accurate and representative of modern norms.

IV. A PROPOSED SAFE HARBOR FRAMEWORK

While many legal scholars and practitioners recognize the importance of using copyrighted data for data mining, most argue that fair use is enough to

copied version is made available on the same terms to others and acknowledgment of the authors of the Wikipedia article used is included . . . .").

[166] *See* Levendowski, *supra* note 132, at 610.

[167] *Id.*

[168] *Id.*

[169] *Id.* at 615.

[170] *Id.* at 618; *Wikipedia:Who Writes Wikipedia?*, WIKIPEDIA, https://en.wikipedia.org/wiki/Wikipedia:Who_writes_Wikipedia%3F.

[171] *See* Levendowski, *supra* note 132, at 619 ("The English Wikipedia article about New England Patriots tight-end Rob Gronkowski is nearly 4,000 words long and boasts 66 citations. By comparison, Stanleyetta Titus, a revolutionary suffragette and the first woman admitted to the New York state bar, does not even have an article." (footnote omitted)).

[172] *See* Enric Junqué de Fortuny, David Martens & Foster Provost, *Predictive Modeling With Big Data: Is Bigger Really Better?*, 1 BIG DATA 215, 215 (2013) ("This study provides a clear illustration that larger data indeed can be more valuable assets for predictive analytics."); John R. Smith, *IBM Research Releases 'Diversity in Faces' Dataset to Advance Study of Fairness in Facial Recognition Systems*, IBM RSCH. BLOG (Feb. 15, 2019), https://www.ibm.com/blogs/research/2019/01/diversity-in-faces ("The AI systems learn what they're taught, and if they are not taught with robust and diverse datasets, accuracy and fairness could be at risk.").

justify this use in the United States.[173] However, as argued in the previous Part, fair use has its limitations, and uncertainty is still apparent among AI creators.[174] A safe harbor for data mining in U.S. copyright law would provide a layer of legal certainty on top of the broad fair use doctrine. This final Part will outline a proposed framework for a safe harbor for data mining in U.S. copyright law.

A. OTHER TEXT AND DATA MINING (TDM) EXCEPTIONS

As the benefits of AI technologies have become apparent, countries have amended their copyright laws to promote innovation and remain competitive in the AI/ML space.[175] The United States already has comparably strong protections for data miners through the fair use doctrine.[176] However, a general safe harbor for data mining would further the spirit of these protections by clarifying the interaction of copyright and AI for smaller innovators.[177] Other jurisdictions have recently enacted text and data mining (TDM) exceptions that can provide guidance here.

Japan was the first country in the world to update its copyright laws to include an exception for text and data mining.[178] Article 47(7) was introduced in 2009 and authorized TDM by all users for all purposes, whether commercial or non-commercial.[179] In line with Prime Minister Shinzo Abe's objective of promoting AI and Big Data industries, the 2018 Amendment to the Copyright Act later introduced three provisions to clarify the law and remove perceived copyright barriers to AI: (1) Article 30-4 authorizes users to "analyse and understand copyrighted works for machine learning;" (2) Article 47-4 "permits electronic incidental copies of works;" and (3) Article 47-5 "allows the use of copyrighted works for data verification."[180] Japan's three new copyright provisions that specifically reference acts relevant to the machine learning process demonstrate a national commitment to the flourishing of AI industries.

[173] *See, e.g.*, Carroll, *supra* note 78, at 936; Levendowski, *supra* note 132, at 619; U.S. PAT. & TRADEMARK OFF., *supra* note 6, at 26.

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

[175] EUR. ALL. FOR RSCH. EXCELLENCE, THE GLOBAL AI RACE 2 (2018), http://eare.eu/assets/uploads/2018/06/Global-AI-Race.pdf.

[176] *See* U.S. PAT. & TRADEMARK OFF., *supra* note 6, at iv ("Most commenters found that existing fair use law does not require modification, as fair use is a flexible doctrine and is capable of adapting to the use of copyrighted works in the context of AI.").

[177] *See* Menell, *supra* note 40, at 346–48 (discussing fair use discussion).

[178] *See Japan Amends its Copyright Legislation to Meet Future Demands in AI and Big Data*, EUR. ALL. FOR RSCH. EXCELLENCE (Sept. 3, 2018), http://eare.eu/japan-amends-tdm-exception-copyright.

[179] *Id.*

[180] *Id.*

Similarly, the United Kingdom has also enacted an exception for TDM purposes in its copyright laws.[181] However, its exception is markedly narrower than Japan's. Under § 29A of the Copyright, Designs and Patents Act 1988, copies made for TDM analysis do not infringe copyright, provided that the work is carried out "for the sole purpose of research for a non-commercial purpose."[182] The beneficiaries of the exception are also limited to those who have "lawful access" to the work in question.[183] While providing legal certainty for researchers, the U.K. exception leaves out startups and entrepreneurs who aim to commercialize innovative machine learning technologies.

Recognizing the legal uncertainty experienced by researchers who utilize TDM and a need for a harmonized exception among its member states, the European Union recently introduced two exceptions for TDM in Directive 2019/790/EU.[184] Article 3 permits "reproductions and extractions made by research organisations and cultural heritage institutions in order to carry out, for the purposes of scientific research, text and data mining of works or other subject matter to which they have lawful access."[185] Article 4 confers an exception for "reproductions and extractions of lawfully accessible works and other subject matter" for commercial TDM uses but allows rightsholders to opt-out of the exemption.[186] This limitation imposed on commercial TDM uses in Article 4 has been criticized for "effectively creat[ing] and legitimiz[ing] a derivative market for text and data mining, which right holders may wish to control, license[,] or even entirely prohibit."[187] While the TDM exceptions in the Directive recognize the importance of data mining in research and technology, the clear limitations in Article 4 may put commercial AI developers in the European Union at a disadvantage.

Not to be left behind, other countries are also considering changes to their copyright law to provide legal certainty and encourage innovation and research in data mining.[188] Australia is considering several TDM-related exceptions as part of an overall fair use exception.[189] Canadian scholars have advocated for a

[181] Copyright, Designs and Patents Act 1988, c. 48, § 29A (UK).

[182] *Id.*

[183] *Id.*

[184] Council Directive 2019/790, arts. 3–4, 2019 O.J. (L 130) 92, 113–14 (EU).

[185] *Id.*

[186] *Id.*

[187] *See* Bernt Hugenholtz, *The New Copyright Directive: Text and Data Mining (Articles 3 and 4)*, KLUWER COPYRIGHT BLOG (July 24, 2019), http://copyrightblog.kluweriplaw.com/2019/07/24/the-new-copyright-directive-text-and-data-mining-articles-3-and-4.

[188] EUR. ALL. FOR RSCH. EXCELLENCE, *supra* note 175, at 2.

[189] *Id.*

fair use regime or a specific TDM exception for commercial uses.[190] And Singapore is expected to move forward with proposed TDM exceptions enabling broad uses for both commercial and non-commercial contexts.[191] As countries amend and clarify their laws to allow technological flourishing, the United States should also consider revisiting its copyright laws.

B. A U.S. DATA MINING SAFE HARBOR

Lessons can be learned from the implementation of TDM exceptions in other jurisdictions to craft a safe harbor for data mining in the United States. Such a safe harbor should (1) define "data mining" broadly; (2) clearly allow reproductions, derivations, and dataset-sharing for data mining purposes; (3) allow non-commercial and commercial uses to the same extent; and (4) be limited to functional, non-expressive uses of data mining.

A safe harbor should broadly define "data mining" to cover a wide range of pattern extraction techniques, not limited to AI/ML. One potential definition is proposed by Jean-Paul Triaille: "The automated processing of digital materials, which may include texts, data, sounds, images or other elements, or a combination of these, in order to uncover new knowledge or insights" (although he uses the term "data analysis" instead of "data mining").[192] While other jurisdictions refer to "text and data mining," "text" is redundant, as it is encapsulated in "data."[193] In crafting the safe harbor, what matters is not the exact technique through which patterns are extracted, but the automated process of gaining functional insights from expressive data. While this Note specifically considers data mining in a machine learning and artificial intelligence context, using a broad definition of data mining in a copyright safe harbor would not foreclose future technologies that employ other techniques for pattern extraction.

A safe harbor should also be specific enough in the rights that are conferred to data miners to provide legal clarity. Given the intimidating U.S. copyright remedial regime and the unpredictable fair use doctrine, smaller innovators in the United States would benefit from an exception that clearly spells out allowed uses.[194] Here, lessons can be drawn from Japan's recent amendments to clarify its copyright exception, which provide an innovation-friendly

[190] *Id.*

[191] *Id.*

[192] JEAN-PAUL TRIAILLE, JÉRÔME DE MEEÛS D'ARGENTEUIL & AMÉLIE DE FRANCQUEN, STUDY ON THE LEGAL FRAMEWORK OF TEXT AND DATA MINING (TDM) 17 (2014), https://www.fosteropenscience.eu/sites/default/files/pdf/3476.pdf.

[193] *See id.* at 8–9.

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

and clear implementation of an exception.[195] It should be clearly enumerated that reproductions and derivations (for labeling and annotation) of works are permitted in a data mining context. Further, the secure storage, retention, and sharing of datasets should also be permitted for verification purposes. Allowing researchers and regulators the ability to scrutinize how data is being used would promote fairer, safer, and higher-quality AI products.[196]

Policymakers should also be careful to not impose limitations on commercial uses, such as opt-out mechanisms, that could unduly stifle innovation contrary to the expressed policy of copyright law in the United States.[197] Unlike the exceptions of the United Kingdom and European Union, commercial uses should be permitted to the same extent that research and non-profit uses are allowed. The U.S. fair use doctrine already includes a copyright exception for research uses and considers nonprofit uses favorably.[198] Thus, the proposed safe harbor would need to specifically free up commercial uses of data mining. Rightsholders should not be allowed to opt-out of uses of their works for commercial data mining purposes because doing so would legitimize a licensing market for machine learning training data that could prompt further legal confusion, put smaller innovators at a competitive disadvantage, and lead to lower quality, biased AI systems.[199]

Finally, and most importantly, the safe harbor should be limited to functional, non-expressive uses of data mining. Some emerging AI technologies use expressive data to generate expressive works and are not covered within the scope of this Note.[200] Because expressive AI-generated works could compete with and effectively replace original works, it would be more difficult to justify a categorical safe harbor for these works without unduly encumbering

[195] EUR. ALL. FOR RSCH. EXCELLENCE, *supra* note 178.

[196] *See* Levendowski, *supra* note 132, at 605.

[197] *See* Christophe Geiger, Giancarlo Frosio & Oleksandr Bulayenko, *Text and Data Mining: Articles 3 and 4 of the Directive 2019/790/EU* 36 (Ctr. for Int'l Intell. Prop. Stud., Research Paper No. 2019-08, 2019); Authors Guild v. Google, Inc., 804 F.3d 202, 212 (2d Cir. 2015) ("The ultimate goal of copyright is to expand public knowledge and understanding . . . . [T]he ultimate, primary intended beneficiary [of copyright] is the public . . . .").

[198] 17 U.S.C. § 107.

[199] *See* Hugenholtz, *supra* note 187; Levendowski, *supra* note 132, at 579.

[200] *See, e.g.*, Chris Baraniuk, *Computer Paints 'New Rembrandt' After Old Works Analysis*, BBC NEWS (Apr. 6, 2016), https://www.bbc.com/news/technology-35977315; Tom May, *AI 'Artist' Creates New Work Inspired by Jean-Michel Basquiat*, CREATIVE BOOM (Sept. 14, 2020), https://www.creativeboom.com/inspiration/ai-artist-resurrects-the-iconic-jean-michel-basquiat-to-mark-32-years-since-his-death-; Steven Poole, *The Rise of Robot Authors: Is the Writing on the Wall for Human Novelists?*, THE GUARDIAN (Mar. 25, 2019, 4:00 AM), https://www.theguardian.com/books/2019/mar/25/the-rise-of-robot-authors-is-the-writing-on-the-wall-for-human-novelists.

authors and creators. Instead, the fair use doctrine should be applied on a case-by-case basis for expressive uses of data mining.

In sum, a data mining safe harbor should (1) define "data mining" broadly; (2) clearly allow reproductions, derivations, and dataset-sharing for data mining purposes; (3) allow non-commercial and commercial uses to the same extent; and (4) be limited to functional, non-expressive uses of data mining.

V. CONCLUSION

As the United States strives to maintain its dominance in artificial intelligence,[201] it must consider how existing laws enable or stifle technological progress. Copyright law presents a potential barrier for AI growth when machine learning models are trained using expressive data.

Fundamentally, data mining is not copyright infringement. While the fair use doctrine provides a degree of flexibility in U.S. copyright law, certainty in the law is desirable for smaller actors. Practically, the fair use doctrine is unpredictable and has been stretched beyond its limits in new technological contexts.

Establishing a clear right to use copyrighted materials to train functional machine learning models is consistent with the goals of copyright law and would ultimately remove barriers to innovation. Rather than rely on courts to rule on important issues of technology policy, Congress should be more proactive in enacting a data mining safe harbor. A safe harbor would make legislatively clear what should already be clear given the scope of copyright law: data mining is not infringement at all.

[201] *See* Exec. Order No. 13,859, 84 Fed. Reg. 3967 (Feb. 14, 2019).