BU School of Law

BU/MIT Student Innovations Law Clinic
765 Commonwealth Avenue
Boston, Massachusetts 02215
(617) 353-3131

Before the

United States Copyright Office

In the Matter of
Request for Comments Regarding Artificial Intelligence and Copyright

Docket No. 2023-6; COLC-2023-0006

Comments of
Robert Mahari, Shayne Longpre,
Kurt Bollacker, Niklas Muennighoff, Nathan Khazam, and Sandy Pentland

Lisette Donewald
Alan Polozov
Clinical Students
Ari Lipsitz
Counsel for Commenters

October 30, 2023

About the Commenters

The Commenters are researchers who recently convened the Data Provenance Initiative, a multi-disciplinary effort between legal and machine learning experts to systematically audit and trace over 1,800 popular machine learning dataset licenses.[1] Robert Mahari received his J.D. from Harvard Law School and is currently a PhD candidate at the MIT Media Lab.[2] He studies the intersection of machine learning and the practice of law. Shayne Longpre is an applied machine learning scientist, as well as a PhD candidate at the MIT Media Lab. Kurt Bollacker, Niklas Muennighoff, and Nathan Khazam are computer scientists and co-authors of the Data Provenance Initiative study. Sandy Pentland is a computer scientist, the director of MIT's Human Dynamics Laboratory and advisor on the study. The Commenters believe that the Copyright Office would benefit from the findings associated with their scholarship, which identifies severe deficiencies in the licensing ecosystem regarding supervised datasets.

I. Summary of Argument

The purpose of this Comment is to address some of the licensing issues that stem from using supervised datasets to train generative AI. Scholars have paid much attention to the copying of raw data to train and develop machine learning models.[3] Many have argued that such use of raw data, derived either directly from the internet or from a dataset, is protected under fair use such that the owners of the original work may not be successful in a claim for copyright infringement.[4] We refer to such compilations of data derived from another source, and repurposed for machine learning, as *unsupervised datasets*. Less attention, however, has been paid to *supervised datasets*, which we define as datasets containing data created for the sole purpose of training machine learning models (mainly for finetuning and alignment). Supervised datasets may likely contain copyrightable contributions from the dataset creators in the form of annotations.[5] To the extent that dataset creators likely have copyright interests in their supervised

[1] Shayne Longpre et al., *The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI* (2023) [hereinafter *Data Provenance*], *available at* https://www.dataprovenance.org/paper.pdf.
[2] All institutional names are for identification purposes only.
[3] *See, e.g.*, Peter Henderson et al., *Foundation Models and Fair Use*, at 5 (2023) (unpublished manuscript) ("[S]ome legal scholars believe that fair use covers most types of model training where the resulting model functions differently than the input data, particularly when the model targets a different economic market[.]"), *available at* https://arxiv.org/pdf/2303.15715.pdf; Mark A. Lemley & Bryan Casey, *Fair Learning*, 99 TEX. L. REV. 743, 748 (2020) (arguing that "an ML system's use of the data often is transformative . . . because even though it doesn't change the underlying work, it changes the purpose for which the work is used"), *available at* https://texaslawreview.org/fair-learning/; Benjamin L.W. Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 COLUM. J.L. & ARTS 45, 59 (2017), *available at* https://doi.org/10.7916/jla.v41i1.2036.
[4] Henderson et al., *supra* note 3; Lemley & Casey, *supra* note 3; Sobel, *supra* note 3; U.S. Patent & Trademark Office, *Public Views on Artificial Intelligence and Intellectual Property Policy* 26 (2020) ("Most commenters found that existing law does 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.").
[5] *Data Provenance*, *supra* note 1, at 14.

datasets, model developers must either rely on fair use or a license in order to avoid infringing the work of dataset creators.[6]

However, we argue that the unauthorized use of supervised datasets is unlikely to be protected by fair use. Whereas the use of unsupervised data for training machine learning is distinct from the original purpose of the unsupervised data, the unauthorized use of supervised datasets for training machine learning is identical to its original purpose.[7] Fair use would therefore likely not apply to the annotations, labels, and curated comments in supervised datasets. For this reason, having a valid license to a supervised dataset is perhaps particularly critical.

Unfortunately, our recent research has found that the licenses attached to publicly available supervised datasets are often imprecise, inaccurate, or missing altogether.[8] Model developers may be exposing themselves to unknown amounts of liability.[9] We argue that this is a problem that needs to be addressed and propose a tool that might serve as a launching point for ensuring license transparency.

The subject matter addressed in this Comment implicates questions 3, 6, 6.1, 6.2, 8, 8.1, 9.1 and 10 within the Copyright Office's Notice of Inquiry.

II. The copyright interests in unsupervised and supervised datasets may be different.

We distinguish unsupervised datasets from supervised datasets because the copyright interests between the two may be different.[10] As shown in Table 1, unsupervised data that is scraped from the web directly and compiled into a dataset involves the copyright interest of the original author of the underlying data in the dataset and the much lesser copyright interest of the dataset creator.[11] In contrast, supervised data also contains the added expressive content of annotations, which were created for the sole purpose of machine learning.[12] A summary of this Comment's nomenclature and argument is below in Table 1:

[6] *See* Sony Corp. of Am. v. Universal City Studios, Inc., 464 U.S. 417, 433 (1984) (Anyone who is authorized by the copyright owner to use the copyrighted work in a way specified in [the Copyright Act] ... is not an infringer of the copyright with respect to such use.").
[7] *Data Provenance*, *supra* note 1, at 15 ("In stark contrast to the copyrighted content that is scraped from the web, supervised datasets were created for the sole purpose of furthering machine learning.").
[8] *Id.* at 8.
[9] *Id.*
[10] *Id.* at 15.
[11] *See, e.g.*, Thomson Reuters Enter. Centre GMBH et al. v. Ross Intelligence, Inc., No. 1:20-cv-613-SB, 2023 WL 6210901 (D. Del. September 25, 2023) (noting that the plaintiff asserted a copyright in the underlying data); Feist Publ'ns, Inc. v. Rural Tel. Serv. Co., 499 U.S. 340, 349 (1991) (noting that the copyright in a factual compilation is "thin").
[12] *Data Provenance*, *supra* note 1, at 15.

Table 1: Comparison of unsupervised and supervised data				
Data Type	**How acquired**	**Potential copyright interests**	**Fair use analysis**	**Examples**
Unsupervised data	Scraped by model developer	Third parties: Underlying content Model developer: "Thin" compilation	Distinct purposes between underlying content and training; fair use likely to apply (see Section 3.a).	Writing Prompts (database of writing prompts and stories scraped on Reddit and used to build a story generation model)[13]
	Via third-party dataset containing minimal modifications or additional expressive content	Third parties: Underlying content and "thin" compilation		WIT3 (TED Talk transcriptions)[14] Ubuntu Dialogue Corpus (chat logs from technical customer support chats for Ubuntu-related problems)[15]
Supervised data	Created by dataset creator	Dataset creator: Expressive annotations and "thin" compilation	Identical purposes between annotations and training; fair use unlikely to apply (see Section 3.b)	The Winograd Schema Challenge (pairs of sentences that differ only in one or two words and that contain a referential ambiguity that is resolved in opposite directions in the two sentences)[16] Deal or No Deal (Dialogs between two people hired by the dataset creator to engage in a negotiation task)[17]

Table 1: Comparison of unsupervised and supervised data				
	Via third-party dataset containing extensive modifications or additional expressive content	Third parties: Underlying content (may be minimal) Dataset creator: Expressive annotations and "thin" compilation		SQuAD (annotated Wikipedia articles for reading comprehension)[18] Stanford Sentiment Treebank (sentences from movie reviews on Rotten Tomatoes labeled as positive or negative by human annotators)[19]

a. Unsupervised data may contain copyright interests in the underlying raw data as well as the compilation of such data.

Definition and usage. Unsupervised datasets typically are unlabeled collections of raw data.[20] Raw data includes data created when users interact with internet platforms, which can be easily accessible for developing generative AI systems.[21] Such data can include individuals

[13] Angela Fan et al., *Hierarchical Neural Story Generation*, *in* PROCEEDINGS OF THE 56TH MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS 889 (2018), *available at* https://aclanthology.org/P18-1082.pdf.
[14] Mauro Cettolo et al., *WIT3: Web Inventory of Transcribed and Translated Talks*, in PROCEEDINGS OF THE 16TH EAMT CONFERENCE 261 (2012), *available at* https://aclanthology.org/2012.eamt-1.60.pdf.
[15] Ryan Lowe et al., *The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems*, *in* PROCEEDINGS OF THE 16TH ANNUAL MEETING OF THE SPECIAL INTEREST GROUP ON DISCOURSE AND DIALOGUE 285 (2015), *available at* https://aclanthology.org/W15-4640.pdf.
[16] Hector J. Levesque et al., *The Winograd Schema Challenge*, *in* PROCEEDINGS OF THE THIRTEENTH INTERNATIONAL CONFERENCE ON PRINCIPLES OF KNOWLEDGE REPRESENTATION AND REASONING 552 (2012), *available at* https://cdn.aaai.org/ocs/4492/4492-21843-1-PB.pdf.
[17] Mike Lewis et al., *Deal or No Deal? End-to-End Learning for Negotiation Dialogues*, *in* PROCEEDINGS OF THE 2017 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING 2443 (2017), *available at* https://aclanthology.org/D17-1259.pdf.
[18] Pranav Rajpurkar et al., *SQuAD: 100,000+ Questions for Machine Comprehension of Text*, *in* PROCEEDINGS OF THE 2016 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING 2383 (2016), *available at* https://aclanthology.org/D16-1264.pdf.
[19] Richard Socher et al., *Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank*, *in* PROCEEDINGS OF THE 2013 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING 1631 (2013), *available at* https://aclanthology.org/D13-1170.pdf.
[20] *See, e.g.*, Alec Radford et al., *Improving Language Understanding by Generative Pre-Training*, at 4 (2018) (unpublished) (discussing unsupervised dataset of over 7,000 unpublished books across a variety of genres), *available at* https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.
[21] *See* Katherine Lee et al., *Talkin' 'Bout AI Generation: Copyright and the Generative AI Supply Chain*, at 30 (unpublished) (last revised Sept. 21, 2023), *available at* https://ssrn.com/abstract=4523551.

sharing artworks (e.g. DeviantArt), writing product reviews (e.g. Amazon Reviews), or engaging in online discussion forums (e.g. Reddit).[22] Generative AI models use unsupervised datasets to identify patterns in the raw data, without specific features that have been labeled by a human.[23] By focusing on identifying patterns in unsupervised datasets, machine learning models can extract or mimic patterns that a human might not necessarily discern, but which still express the essential features of a group of raw data.[24]

An example of an unsupervised dataset is the Web Inventory of Transcribed and Translated Talks (WIT3). This dataset was created to offer access to a collection of transcribed and translated TED Talks, which are distributed by the TED website under a Creative Commons license prohibiting commercial use, forbidding derivative works, and requiring attribution.[25] Due to its size, variety of topics, and covered languages, this unsupervised dataset is an excellent resource for the machine translation research community.[26]

Copyright interests. To the extent there are copyright interests in the raw data used in unsupervised datasets, they may be held by the creator of that raw data or another third party (such as the platform on which the data was posted). However, an unsupervised dataset is not necessarily entirely comprised of individual examples of copyrighted data.[27] For example, certain material may not satisfy the originality requirement.[28] If the author of an individual example of raw data did not contribute the necessary modicum of creativity to the work, the raw data example is not original and is not protected by copyright.[29] For example, recordings of birds used to train birdsong-recognition AI models may lack human contributions with the requisite modicum of creativity, and so the individuals who recorded the birdsongs may not be able to claim copyright.[30] Other material may be more expressive, such as illustrations in an unsupervised dataset used to train image recognition models.[31]

In addition to the copyright interest in the underlying raw data held by the creator of the original work or platform on which the work is posted, the creator of the unsupervised dataset may have a copyright interest in the dataset itself, to the extent the dataset contains an original selection and arrangement of the underlying data.[32] However, this protection would be thin, especially if the selection and arrangement is not very original.[33]

---

[22] *Data Provenance*, *supra* note 1, at 14.
[23] Sobel, *supra* note 3, at 59. "Unsupervised learning, by contrast, apprehends patterns in data without being prompted with a particular kind of output; it just uncovers 'interesting structure.'"
[24] *Id.*
[25] *See TED Talks Usage Policy* (last accessed Oct. 29, 2023), https://www.ted.com/about/our-organization/our-policies-terms/ted-talks-usage-policy.
[26] Cettelo et al., *supra* note 14, at 1.
[27] *Data Provenance*, *supra* note 1, at 14.
[28] 17 U.S.C. §102(a); *see Feist*, 499 U.S. at 346 (stating that originality requires "independent creation plus a modicum of creativity").
[29] *See Feist*, 499 U.S. at 358.
[30] Lee et al., *supra* note 21,21 at 52.
[31] *Id.*
[32] *See Feist*, 499 U.S. at 348; *see also* 17 U.S.C. § 101.
[33] *See, e.g.*, Compendium of U.S. Copyright Practices (Third) § 312.2 (finding unregistrable compilations "consisting of all the elements from a particular set of data" or "containing only two or three elements"); Experian Info. Solutions v. Nationwide Mktg., 893 F.3d 1176, 1187 (9th Cir. 2018) (finding compilations of factual credit data involved at least minimal creativity, but afforded "thin protection" requiring "substantial verbatim copying").

b. Supervised datasets may additionally contain a copyrightable interest in any creative annotations authored by the dataset creator(s).

Definition and usage. While unsupervised datasets are useful in training models, they tend to be insufficient for eliciting high performance from generative AI models.[34] Consequently, supervised dataset creators set out to generate custom datasets to improve the performance of machine learning models on specific tasks or fine-tune them.[35] These datasets may be created out of whole cloth, for example by asking experts to write logical statements, or they may build on raw data, for example by asking annotators to extract question-answer pairs from articles.[36] In either case, supervised dataset creators make extensive curatorial choices and add customized expression in the form of labels and annotations layered on top of raw data.[37] These labels and annotations are typically human-made, although increasingly supervised data is created via large language models.[38] Supervised datasets are also not necessarily made by a single actor. Rather, they can be created by multiple entities through several stages of scraping and annotation of the raw data.[39]

A prototypical example of a supervised dataset is SQuAD, built to train algorithms on reading comprehension.[40] To create the dataset, dataset creators extracted paragraph-long excerpts from 539 Wikipedia articles, and enlisted humans to generate over 100,000 questions answered by the excerpts. For example:

- **Wikipedia excerpt**: *In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity.*
- **Worker-generated question**: *What causes precipitation to fall?*
- **Answer**: Gravity[41]

Copyright interests. As with unsupervised datasets, supervised datasets also contain varying amounts of copyrighted material.[42] Like unsupervised datasets, supervised datasets may consist of copyrightable interests in the underlying raw data held by the data creator/platform, as well as in any original selection and arrangement of the dataset as a compilation of that data held by the dataset creator. (In practice, we observe that supervised datasets may tend to include far less third-party data than unsupervised datasets. For example, the SQuAD dataset discussed above contains limited excerpts of only 539 Wikipedia articles.)

However, we argue that there is an important distinction between supervised and unsupervised datasets for the purposes of copyright analysis. ***First***, the annotations and labels—many (but not all) of which contain original human-made authorship—add another layer of expression that may be owned by the dataset creator, beyond any original selection and

[34] Victor Sanh et al., *Multitask Prompted Training Enables Zero-Shot Task Generalization*, at 10 (2022), *available at* https://arxiv.org/abs/2110.08207.
[35] *Data Provenance*, *supra* note 1, at 17.
[36] *See, e.g.*, Rajpurkar et al., *supra* note 18; Levesque et al., *supra* note 16, at 553.
[37] *Data Provenance*, supra note 1, at 14.
[38] *Id.* at 17 (reporting that approximately 12% of popular supervised datasets were annotated using OpenAI).
[39] *Id.*; *compare* Lee et al., *supra* note 2121, at 33 (discussing similar process for unsupervised datasets).
[40] *See* Rajpurkar et al., *supra* note 18.
[41] *Data Provenance*, *supra* note 1, at 14.
[42] *Id.*

arrangement within the dataset itself.[43] Some supervised datasets may be highly expressive, such as a supervised dataset containing negotiation dialogues.[44] ***Second***, as we discuss further below, whereas unsupervised datasets contain raw data created for a myriad of purposes unrelated to machine learning, the annotations within a supervised dataset were purpose-built for training machine learning models.[45]

Assuming the copying of supervised datasets implicates the exclusive rights of copyright, an actor who wants to use a supervised dataset needs to rely on permissions from the supervised dataset creator, through a license, or the actor needs to rely on an exception to infringement through fair use.[46]

III. Because supervised datasets were created for the sole purpose of training machine learning, they are less likely to be fair use.

Courts have not conclusively provided an answer to whether, or when, fair use applies to data for machine learning.[47] Accordingly, one of the purposes of this Comment is to draw a distinction between how fair use likely applies to using unsupervised data compared to supervised datasets for the purposes of training generative AI models. (This Comment does not take a position on the extent to which the generation of expressive *output* by such models may or may not be fair use.[48])

The fair use balancing test considers four factors which are: (1) the purpose and character of the use, and whether the use is for a commercial or nonprofit, educational purposes; (2) the nature of the copyrighted work; (3) the amount and substantiality of the portion copied in relation to the whole; and (4) the effect of the use upon the potential market for or value of the copyrighted work.[49] These factors are given varying weight in the analysis depending on context.[50]

a. Fair use may apply to the use of unsupervised data to train AI models.

As outlined in Table 1 above, and as discussed further above, there are two broad categories of unsupervised data usage. Model developers may directly acquire raw data, or else rely on third-party datasets that have been compiled for purposes of training machine learning. In our view, there is little meaningful distinction between these two cases. Although a dataset creator who scraped raw unsupervised data may have a copyright in the compilation of the

---

[43] *Id.*
[44] *See* Lewis et al., *supra* note 17.
[45] *Data Provenance*, *supra* note 1, at 15.
[46] *See Sony Corp.*, 464 U.S. at 433 ("Anyone who is authorized by the copyright owner to use the copyrighted work in a way specified in [the Copyright Act] ... is not an infringer of the copyright with respect to such use.").
[47] *See, e.g.*, *Thomson-Reuters*, 2023 WL 6210901, at *8 (denying cross-motions for summary judgment on fair use defense as applied to use of competitor's dataset for asserted purpose of machine learning, stating the "precise nature" of defendant's actions must be decided by jury).
[48] *See generally* Sobel, *supra* note 3, at 61-65 (distinguishing between dataset creation, model training, and model output).
[49] 17 U.S.C. § 107.
[50] *See* Andy Warhol Foundation for the Visual Arts, Inc. v. Goldsmith, 143 S. Ct. 1258, 1274 (2023) ("The Copyright Act's fair use provision . . . 'set[s] forth general principles, the application of which requires judicial balancing, depending upon relevant circumstances.'") (quoting Google LLC v. Oracle Am., Inc., 141 S. Ct. 1183, 1197 (2021)).

resulting dataset, this will necessarily be quite thin.[51] For this reason, in practice, litigation related to the unauthorized acquisition and use of training data for machine learning purposes has emphasized the copyright interests of the creators of the underlying raw data, rather than the limited copyright interests of the dataset creators.[52] Accordingly, we will focus the fair use analysis regarding unsupervised data on the *content* of such data, rather than on any copyright interest in the *dataset*. (However, as discussed below, we believe the reverse holds for *supervised* data, the value of which derives from the content of the copyrightable annotations made by the dataset creator.)

Although fair use is a case-by-case determination, we believe that the principles of fair use permit the use of unsupervised datasets to train generative AI models. This is provided that the underlying works are significantly "transformed" into model weights, only a small amount of training data is retained by the trained model, model training is designed to only glean generally generalizable insights from the training data, and the trained model does not have a strong effect on the economic success of the works in the training data.[53]

Factor One. With respect to the first factor, the Supreme Court's recent decision in *Warhol Foundation v. Goldsmith* can provide guidance. There, the Court held that the first factor likely disfavors fair use where (1) the secondary use shares "the same or highly similar purposes," (2) "is of a commercial nature," and (3) "some other justification for copying" is absent.[54] In assessing "purpose," the majority focused its inquiry on the "environment" or "objectives" of the secondary use, rather than on any intrinsic "meaning or message."[55]

As to unsupervised datasets, the secondary use is distinct from the purpose of the underlying raw data, which likely favors fair use under the first factor.[56] This underlying content was created for any number of purposes depending on what the dataset consists of and depending on the original context.

One example might be a dataset collection of fairy tales. The purpose of the underlying data—the fairy tales—may be to be sold as stories. In contrast, when those fairy tales are used to train a generative AI model, they are used for an entirely different purpose: for example, to extract generalizable insights and patterns from language to facilitate the generation of realistic text.[57] On its own, the unsupervised dataset is unlikely to run the risk of substituting the fairy tales in their original context. Thus, the purpose of the unsupervised dataset is a distinct secondary use that does not share the "same or highly similar purpose" as the underlying raw data.[58]

---

[51] *See Feist*, 499 U.S. at 349.
[52] *See, e.g.*, Authors Guild v. OpenAI Inc., No. 1:23-cv0-08292 (S.D.N.Y. Sept. 19, 2023); Silverman et al. v. OpenAI, Inc. et al., No. 3:23-cv-03416 (N.D. Cal. Jul. 7, 2023); Getty Images (US), Inc. v. Stability AI, Inc., No. 1:23-cv-00135 (D. Del. Feb. 3, 2023); Andersen v. Stability AI Ltd., No. 3:23-cv-00201 (N.D. Cal. Jan. 13, 2023).
[53] *Data Provenance*, *supra* note 1, at 16. This Comment does not take a position on the extent to which model weights are "transformed" for purposes of assessing whether the resulting model constitutes an infringing derivative work or a valid transformative use.
[54] *See Warhol*, 143 S. Ct. at 1277.
[55] *See id.* at 1279, 1282.
[56] *See* Lemley & Casey, *supra* note 3
[57] *See Data Provenance*, *supra* note 1, at 16.
[58] *See Warhol*, 143 S. Ct. at 1277.

The remaining considerations of the first factor also support fair use. With respect to commerciality, the secondary use of the underlying raw data is likely commercial in nature, considering that there is a thriving market for training data.[59] However, it is important to note that the context in which unsupervised datasets are used is distinct and different (i.e., machine learning) from the context in which the works that comprise the underlying raw data would be licensed.[60] In the example of an unsupervised dataset collection of fairy tales, the context for the underlying fairytales would be one that seeks to facilitate people's engagement with the stories' expression.[61] In contrast, unsupervised datasets are used to create new systems and products.[62] Even though the secondary use of the underlying raw data is likely commercial in nature, this consideration as part of the first factor analysis is unlikely to weigh against a finding of fair use, as it did in *Google*.

Finally, courts may find a compelling justification to use the underlying raw data for unsupervised datasets. Machine learning is chiefly a predictive technology, which learns by analyzing vast amounts of input data to discern patterns without human intervention.[63] Early AI, like the "expert system," relied on analyzing hard-coded knowledge baked in by human designers, rather than analyzing troves of unsupervised data.[64] Early computer programs like the expert system were capable of generating content that could resemble, though not rival, human expression.[65] Today, because of new machine learning techniques, including training AI models on unsupervised data, AI models have far greater capabilities beyond what they could do when they relied on a small knowledge base of facts, rules derived from those facts, and an inference engine for reaching conclusions.[66] For example, these new machine learning techniques enable machines to identify and mimic features that distinguish sensory data, even when those features are not qualities that humans can easily express or represent.[67] Machine learning is also capable of powerful reasoning abilities in large part due to the diversity and richness of ever larger training datasets, including pre-training with unsupervised data.[68] The compelling justification for using underlying raw data is that machine learning models are designed to work better when trained on a broad range of content.[69]

Factor Two. The second factor focuses on the extent to which the copyrighted work is close to the "core of copyright," including considerations about whether it is primarily expressive or functional in nature, and whether it has been published.[70] Unsupervised datasets can vary from

[59] *See* Sobel, *supra* note 3, at 76.
[60] *See Google*, 141 S. Ct. at 1206-07; *see also Warhol*, 143 S. Ct. at 1277 n.8 (characterizing *Google*'s opinion finding factor one supported copying for a "new system created for new products").
[61] *See* Sobel, *supra* note 3, at 57.
[62] *Warhol*, 143 S. Ct. at 1277 n.8 (characterizing *Google*'s opinion finding factor one supported copying for a "new system created for new products").
[63] Sobel, *supra* note 3, at 58.
[64] *Id.*
[65] *Id.*
[66] *Id.*
[67] *Id*. at 60.
[68] *Data Provenance*, *supra* note 1, at 1; *see generally* Katherine Lee et al., *The Devil is in the Training Data*, in AI AND LAW: THE NEXT GENERATION (2023) [hereinafter *The Devil is in the Training Data*].
[69] *See* Sobel, *supra* note 3, at 58; *compare Google*, 141 S. Ct. at 1203-04 (reviewing how reimplementation of API interfaces "can further the development of computer programs"); *see generally The Devil is in the Training Data*, *supra* note 68.
[70] *Google*, 141 S. Ct. at 1202.

being expressive to primarily informational, as discussed above.[71] If the raw data is primarily informational, its use is more likely to be fair use.[72] Most of the raw data will typically be published within the meaning of copyright law, which may also tilt this factor in favor of fair use.[73] For use of raw data that is unpublished within the meaning of copyright law, this factor would likely disfavor fair use.[74]

Factor Three. Much like the second factor, analysis of the third factor varies depending on the amount and substantiality that the unsupervised dataset's raw data has copied from original works.[75] If the raw data contains relatively small portions from original works, then this factor may weigh more in favor of fair use (although there is no set proportion).[76] Typically, however, an unsupervised dataset "copies complete works verbatim."[77] Thus, at least the initial use of such datasets may result in an initial complete copy of the underlying works.[78] Nonetheless, this complete copying may be justifiable to the extent necessary for the transformative purpose of training an AI model.[79] Moreover, over the course of training a model, the model is unlikely to retain a full "copy" of the initial dataset. Conversely, if the raw data contains more of the original works than is necessary for its secondary use, then this factor may weigh against a finding of fair use.[80]

Factor Four. The fourth factor considers the "effect" of the copying in the "market for or value of the copyrighted work."[81] Only harms "cognizable under the Copyright Act" are considered to disfavor fair use.[82] There is certainly a large market for licensing unsupervised training data. However, as discussed above, this serves a different purpose from the underlying works. (Again, this analysis is limited to datasets for *training* AI models—to the extent there is an *output* generated from those models that competes with the original works, there may be market harm.) Therefore, the fourth factor likely weighs in favor of a finding of fair use, provided that the trained model does not have a strong effect on the economic success of the works in the training data.[83]

Overall, in light of the distinct purposes and distinct markets between the underlying unsupervised data and the machine learning use, we believe that model developers should anticipate that courts may find certain uses of unsupervised datasets to constitute fair use.[84]

---

[71] Lee et al., *supra* note 2121, at 96.
[72] *Id.*
[73] *Id.*
[74] *Id.*; *see also Harper & Row Publishers, Inc. v. Nation Enters.*, 471 U.S. 539, 564 (1985) ("A use that so clearly infringes the copyright holder's interests in confidentiality and creative control is difficult to characterize as 'fair.'").
[75] Henderson et al., *supra* note 3, at 6.
[76] *Id.*
[77] Lee et al., *supra* note 2121, at 102.
[78] Sobel, *supra* note 3, at 62 ("Once an input dataset has been compiled, it may be copied, emulated, and re-copied thousands of times during the learning process.").
[79] *See Google*, 141 S. Ct. at 1205 ("The 'substantiality' factor will generally weigh in favor of fair use where, as here, the amount of copying was tethered to a valid, and transformative, purpose.").
[80] Lee et al., *supra* note 2121, at 96.
[81] 17 U.S.C. § 107(4).
[82] *Google*, 141 S. Ct. at 1206 (citing Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 591-92 (1994)).
[83] *Id.*
[84] Henderson et al., *supra* note 3, at 5.

b. In contrast, fair use is less likely to apply to the use of supervised datasets created for the sole purpose of training machine learning models.

In contrast, the use of supervised datasets is less likely to be fair use. Unlike unsupervised data, the annotations encoded within supervised data were created by dataset creators for the sole purpose of training machine learning models. Accordingly, a dataset creator may have a claim against an unauthorized use of a supervised dataset.

Factor One. Under *Warhol*, the first factor as applied to the copying of supervised datasets likely weighs against a finding of fair use. The underlying content of supervised datasets contains labels, annotations, or other expressive content specifically created to instruct generative AI.[85] It is this additional content, which was specifically created to train generative AI, that distinguishes unsupervised from supervised datasets.[86] Here, if another model developer wished to use a supervised dataset to train her generative AI model, the purpose of her secondary use would be identical to the intended use.

The remaining considerations of factor one similarly disfavor fair use. The secondary use of the underlying content of supervised datasets is likely commercial in nature, since there is a robust market for training data.[87] Both public and proprietary generative models attribute their complex reasoning abilities to the variety of ever-larger training datasets.[88] Model developers are also known to combine and re-package thousands of datasets and web sources.[89] Given that multiple supervised datasets can be amalgamated into collections and those collections are all used for training generative models, the context in which supervised datasets are used (i.e. machine learning) is not distinct and different from the context that licenses the underlying content (i.e. labels, annotations, and other expressive content) of supervised datasets.[90] This consideration as part of the first factor analysis likely also weighs against a finding of fair use.

Finally, while there is likely a compelling justification to copy other supervised datasets, given how powerful the reasoning abilities of machine learning models become the more diverse training datasets they are trained on, this alone may not be enough to favor a finding of fair use under the first factor.[91]

Factor Two. The nature of the annotations within supervised datasets may be expressive, since dataset creators make curatorial choices through labeling and annotation of the underlying content, in addition to choices in the selection and arrangement of the underlying content.[92] The more options that are available for annotation, the more likely the copied material is expressive. Since all supervised datasets embody at least some curatorial choices, all supervised datasets are

[85] Sobel, *supra* note 3, at 59.
[86] *Id.*
[87] *Id.* at 76.
[88] *Data Provenance*, *supra* note 1, at 1.
[89] *Id.*
[90] *Id.* at 16; *see also Warhol*, 143 S. Ct. at 1277 n.8 (characterizing *Google*'s opinion finding factor one supported copying for a "new system created for new products").
[91] *See Warhol*, 143 S. Ct. at 86 (noting that use of copyrighted work merely to convey a new meaning or message is not a compelling justification).
[92] *Data Provenance*, *supra* note 1, at 14; Lee et al., *supra* note 21 at 33-34.

expressive to some extent.[93] Thus, given the expressiveness of the underlying content of supervised datasets, the second factor likely weighs against a finding of fair use.[94]

Factor Three. As with unsupervised datasets, an unauthorized use of a supervised dataset will necessarily copy the totality of the dataset. However, whereas the copying of the underlying data may be justified due to its different purpose, the copyright interest in the supervised dataset includes the expressive annotations copied in full. And unlike unsupervised data, use of supervised data for machine learning is for an identical purpose. Further, given that supervised data is often used at the finetuning or alignment stage, it is more likely that the resulting model will retain the expressive content. Accordingly, the third factor should disfavor fair use in these instances.

Factor Four. As for the fourth factor, the widespread market for licensed training datasets ought to weigh against a finding of fair use.[95] As discussed below, the licensing market for supervised datasets often contains imperfect information—and oftentimes, the open-source licensing terms designed for software may not be well-suited for distribution of datasets.[96] However, the common practice of offering supervised datasets under license ultimately has been driving widespread adoption and use of machine learning among academics and companies.[97] Companies like Scale.ai offer model developers access to diverse datasets by providing annotation services, which further supports the notion that there is a robust market for high quality training data.[98] The structured licensing market makes sense to preserve, especially where the settled expectations of researchers—both dataset creators and model developers—presume that training datasets are used under license.[99]

Consequently, fair use is less likely to apply to supervised datasets than to unsupervised datasets.

IV. Due to these fair use concerns, establishing licenses for supervised datasets is extremely important.

Due to this distinction in how fair use might apply, it is incredibly important that there be transparency in the licenses associated with supervised datasets. An applicable license may be

[93] Lee et al., *supra* note 21, at 3.
[94] *Id.*
[95] *See* Am. Geophysical Union v. Texaco Inc., 60 F.3d 913, 930 (2d Cir. 1994) (considering whether a licensing market is "traditional, reasonable, or likely to be developed").
[96] *Data Provenance*, *supra* note 1, at 16.
[97] Sobel, *supra* note 3, at 83 ("In today's platform economy, value emerges not from the ownership of intellectual property rights in data, but from the ability to make licensed use of large amounts of data.").
[98] Scale AI, *Our mission is to accelerate the development of AI applications*, https://scale.com/about; *see generally* Roger Brown, *Top-6 Data Annotation Companies for AI and Robotics Architects*, MEDIUM (Dec. 18, 2021), *available at* https://cogitotech.medium.com/top-6-data-annotation-companies-for-ai-and-robotics-architects-db843f0727cd.
[99] *Data Provenance*, *supra* note 1, at 16. Commenters recognize this well-known problem of circularity between the existence of a licensing market and the fair use analysis. *See generally* Jennifer E. Rothman, *The Questionable Use of Custom in Intellectual Property*, 93 VA. L. REV. 1899, 1933 (2007). Nonetheless, the distribution of machine learning datasets *has* emerged under a structured paradigm of licensing, and finding fair use may disrupt these settled expectations.

dispositive of an infringement claim.[100] Therefore, for these supervised datasets where fair use is less likely to apply, complying with the scope of a license is critical.

Most supervised datasets are subject to a general license.[101] General licenses allow anyone to use a work for certain purposes, even if a particular user never asked for permission from the copyright owner.[102] An example of this is the Creative Commons CC-BY 4.0 license, which allows certain usage rights to the public such as use for academic purposes in exchange for attribution.[103] A typical supervised dataset publishes annotated data, often for free, subject to a license agreement.[104]

Many supervised datasets are publicly available and subject to various licenses.[105] Popular aggregators like GitHub, Hugging Face, and Papers with Code provide options for the distributors of datasets to report the associated licenses.[106] Most of these dataset licenses are common and recognizable and contain familiar open-source conditions such as attribution and/or share-alike.[107] Furthermore, their allowed uses are commonly commercial, non-commercial, academic, or custom.[108] Additionally, licenses can have any combination of the aforementioned requirements and allowed uses.[109]

Based on our review of the most common supervised datasets that are publicly available, 73% of supervised datasets require attribution and 33% include a share-alike clause.[110] While most licenses are common and recognizable, there are many variants with their own unique requirements, as well as an ample collection of custom licenses.[111]

For example, our review of the major supervised NLP datasets indicated that the most common licenses are the following Creative Commons licenses: CC-BY-SA 4.0, CC-BY 4.0, and CC BY-NC 4.0.[112] CC-BY-SA 4.0 permits licensed works to be used commercially, but requires attribution to indicate the data source, and a share-alike restriction, which subjects any

[100] Oracle Am., Inc. v. Hewlett-Packard Enter. Co., 971 F.3d 1042, 1051 (9th Cir. 2020); *see also Sony*, 464 U.S. at 433 ("Anyone who is authorized by the copyright owner to use the copyrighted work in a way specified in [the Copyright Act] ... is not an infringer of the copyright with respect to such use.").
[101] *Data Provenance*, *supra* note 1, at 7 (Figure 2 of distribution of licenses in supervised datasets).
[102] Lee et al., s*upra* note 21, at 105.
[103] *Id.* at 7; *see also* Jacobsen v. Katzer, 535 F.3d 1373, 1381-82 (Fed. Cir. 2008) (reviewing enforceability of open source license).
[104] General licenses are the dominant model of distributing supervised datasets, but they are not the only means available. Some supervised datasets are granted under specific licenses, which allow a specific named licensee to use the work under the license's terms. *See* Lee et al., *supra* note 2121, at 104. Other supervised datasets are dedicated to the public domain, permitting anyone to use the works without risking copyright infringement. *Id.* at 105; *see* Public Domain Mark 1.0 (2023), https://creativecommons.org/publicdomain/mark/1.0/. Less likely to apply to supervised datasets are implied licenses, which are based on the copyright owner's conduct, indicating consent for particular uses of their work. *Cf.* Lee et al., *supra* note 21,21 at 109.
[105] *Data Provenance, supra* note 1, at 2.
[106] *Id.*
[107] *Id.* at 6.
[108] *Id.*
[109] *Id.* at 7.
[110] *Id.*
[111] *Id.*
[112] *Id.*

derivatives of the data to the same terms as the original license.[113] CC-BY 4.0 similarly permits commercial use in exchange for attribution, but conversely does not require share-alike.[114] Meanwhile, CC BY-NC 4.0 restricts licensed uses to non-commercial or academic purposes.[115]

Importantly, it is rare in practice for a machine learning model to use a single supervised dataset, and often multiple datasets are compiled into collections.[116] This leads to problems when the underlying datasets are subject to conflicting licenses. For example, CC BY-SA 4.0 and CC-BY-NC 4.0 are incompatible; the former requires share-alike and commercial use, whereas the latter does not contain a share-alike provision but restricts to non-commercial use.[117]

Given that different entities can have copyrightable interests in supervised datasets based on their curation and annotation, those entities can subject model developers interested in using their supervised datasets to additional licensing terms.[118] This is why clear and accurate licenses are critical, because broad copyright infringement can occur if model developers are using various supervised datasets when each dataset has its own, potentially conflicting, license.[119]

V. Licensing should be a viable model for distributing supervised datasets, but it is currently broken.

As discussed above, abiding by a given license may be more difficult than model developers expect. Unfortunately, even identifying the governing terms can be impossible. Our research has identified an alarming dearth of accurate provenance information regarding supervised datasets.[120]

Based on our review, many supervised datasets are ambiguously or incorrectly licensed. The licenses for many supervised datasets hosted by popular aggregators are reported incorrectly, and in many cases a more permissive license is listed by the aggregator than by the dataset creators.[121] For example, 66% of the Hugging Face datasets analyzed contained data licensed in a different use category than how they were labeled, and often purported to contain more permissive licensing terms than the author's intended license.[122] Other datasets are missing licenses entirely. Over 70% of licenses on popular dataset sharing sites are "unspecified" regarding their license, leaving model developers to speculate about the risk they may be incurring by using them.[123]

Possible reasons for this absence of accurate provenance could be that aggregators or dataset creators intentionally release supervised datasets without a license, or contributors on

[113] *Id.* at 6; Creative Commons, CC BY-SA 4.0 Deed, *available at* https://creativecommons.org/licenses/by-sa/4.0/deed.en.
[114] *Data Provenance*, *supra* note 1, at 6; Creative Commons, CC BY-4.0 Deed, *available at* https://creativecommons.org/licenses/by/4.0/.
[115] *Data Provenance*, *supra* note 1, at 6; Creative Commons, CC BY-NC 4.0 Deed, *available at* https://creativecommons.org/licenses/by-nc/4.0/.
[116] 1*Data Provenance*, *supra* note 1, at 16.
[117] *See* Creative Commons, *Wiki / CC license compatibility*, https://wiki.creativecommons.org/wiki/Wiki/cc_license_compatibility.
[118] Lee et al., *supra* note 2121, at 110; *Data Provenance*, *supra* note 1, at 13.
[119] *Data Provenance*, *supra* note 1, at 14.
[120] *Id.* at 8.
[121] *Id.*
[122] *Id.* at 2.
[123] *Id.*

these platforms mistake licenses attached to code in the relevant repositories for licenses attached to data.[124]

An additional problem is that, as discussed above, supervised datasets can be composed of numerous underlying datasets. Therefore, even if a particular dataset creator or curator releases a supervised dataset with a chosen license, this does not guarantee that the works within the supervised dataset are properly licensed.[125] All of the above means that many of these supervised datasets are unusable (or harmfully misleading) for risk-averse model developers.[126]

With this context in mind, it is of the utmost importance that supervised datasets have identifiable licenses. This is where we hope our Comment will provide the most value. We believe that if licenses are easier to identify, and easier to reliably attribute, model developers will have a more precise understanding of the biases inherent in their models. Developing reliable attribution may also encourage dataset creators to publish more robust supervised datasets containing data that represents more inclusive perspectives.[127] We therefore stress the importance of having a tool that can identify the underlying licenses attached to supervised datasets.

We have developed one such tool, which we refer to as the Data Provenance Explorer (DPExplorer).[128] It is based on the most extensive audit to date of widely used supervised text datasets in AI.[129] It consists of over 45 of the most popular supervised dataset collections referred to as the Data Provenance Initiative Collection.[130] It annotates dataset identifiers (like name, source URL, and data collection service), characteristics (like topic, number of downloads, and languages the data is in) and provenance (like the data creator, the license, and the license conditions).[131] The information is traced from the original source to the curated collection. Users of the DPExplorer can have more confidence in understanding the licenses attached to the supervised datasets they are using. Specifically, the DPExplorer compiles all self-reported license information, runs a search for explicit data licenses, identifies the license type, categorizes the license, and collects metadata so that model developers can filter for types of licenses and permissibility.[132] Of course, there are limitations, such as that the DPExplorer collects only self-reported licenses. However, it is a meaningful first step toward supporting licensing transparency.

VI. Why does this matter?

The issue of licensing is important for machine learning developers, dataset creators, and the general public. Correctly licensing supervised datasets protects model developers from liability. Because these datasets are predominantly created to help grow the potential of properly

[124] *Id.* at 8.
[125] *Id; see generally The Devil is in the Training Data*, *supra* note 68.
[126] *Data Provenance*, *supra* note 1, at 8.
[127] *Id.* at 17.
[128] *See Data Provenance Explorer*, https://www.dataprovenance.org/.
[129] *Data Provenance*, *supra* note 1, at 2.
[130] *Id.* at 1.
[131] *Id.* at 3.
[132] *Id.* at 6.

trained generative AI models, supervised dataset creators have not yet begun regularly suing for unauthorized usage.[133]

However, provided these datasets are copyrightable and fair use does not apply, it may only be a matter of time before these dataset creators begin regularly filing lawsuits when model developers fail to adhere to licenses governing underlying material.[134] Risk-averse developers are also aware of this possibility, which forces them to avoid using many valuable datasets because they do not have assurances that there are no licenses attached to them.[135] Consequently, if more model developers follow this risk-averse practice, they will only train their generative AI models on a subset of data, which will not be as effective and will hinder progress for generative AI.[136]

For dataset creators, there may be diminished incentive to create these supervised datasets because they know their licenses will not be honored. Although many creators are motivated by the potential capabilities of well-trained generative AI, and they create these supervised datasets to further that purpose, there are other motives as well. At the very least, many creators—especially academic researchers—want attribution or recognition for the work they make. If model developers continue the practice of refusing to honor dataset creators' licenses, it may disincentivize dataset creators from continuing this work.

Finally, a lack of transparency about licenses can have negative impacts for the general public. For risk averse model developers who choose to avoid supervised datasets with unclear licenses, their generative AI models may not receive all the training they need to make accurate judgments.[137] Without diverse, multi-faceted training data, generative AI models may become biased.[138] For example, if a generative AI model has only been trained to recognize English, it may produce worse quality outputs in Spanish. This would bias the model and lead to less accurate outputs. The lack of data provenance can also lead to data leakages between the training set and test data or to the exposure of personally identifying information.[139] Consequently, this can result in license revisions after models are fully trained or deployed, or even lawsuits.[140] And due to the myriad of licenses that exist, startups and less resourced organizations also struggle to navigate responsible training data collection, its legality and ethics.[141]

There are numerous interests at stake here. To ignore the problem of licensing transparency would lead to disincentivizing supervised dataset curators, legal liability for model

---

[133] *Cf.* Lemley & Casey, *supra* note 3, at 746 ("After decades of allowing—or even just plain ignoring—machine copying, copyright owners and courts have begun to loudly and visibly push back against the copyright system's permissive attitude towards machine copying.").
[134] *See supra* note 52.
[135] *Data Provenance*, *supra* note 1, at 8.
[136] Lemley & Casey, *supra* note 33
[137] *Data Provenance*, *supra* note 1, at 2.
[138] *Id.*
[139] *See* Aparna Elangovan et al., *Memorization vs. generalization : Quantifying data leakage in NLP performance evaluation*, *in* PROCEEDINGS OF THE 16TH CONFERENCE OF THE EUROPEAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS 1325 (2021), *available at* https://aclanthology.org/2021.eacl-main.113; Nicholas Carlini et al., *Quantifying memorization across neural language models*, *in* THE ELEVENTH INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS (2022), *available at* https://arxiv.org/abs/2202.07646; Sébastien Bubeck et al., *Sparks of artificial general intelligence: Early experiments with gpt-4* (unpublished) (2023), *available at* https://arxiv.org/abs/2303.12712.
[140] *Data Provenance*, *supra* note 1, at 7.
[141] *Id.*

developers, and credibility and bias issues for generative AI models. A healthy balance needs to be struck, and that can start by making licenses more transparent. As such, we endorse the DPExplorer as a tool for the progress of sustainable and ethical licensing for supervised datasets. We also encourage research on creating dataset specific licenses as opposed to repurposed software licenses. Ultimately, we hope that thoughtful data licensing can be leveraged to promote more responsible, inclusive, and transparent machine learning practices.