Comments of the Center for Democracy & Technology
On the United States Patent and Trademark Office's
Intellectual Property Protection for Artificial Intelligence Innovation

January 10, 2020

The Center for Democracy & Technology (CDT) thanks the U.S. Patent and Trademark Office for the opportunity to comment on these questions. CDT is a non-profit advocacy organization working to preserve individual rights and democratic ideals online, and in existing and new applications of technology. The emergence of machine learning as a path toward artificial intelligence has, and will continue to raise difficult legal and moral questions. Many of these questions may challenge our established systems or force us to reassess their purposes and goals in light of emerging technological capabilities. Yet, not every new technology demands a change in law or policy.

While using the phrase "artificial intelligence" evokes images of super-intelligent robots with human-like capabilities, those concepts have not yet become reality. Instead, much of what we call "AI" now amounts to advanced statistical analysis, and predictions and prognoses based on models derived therefrom. CDT encourages the Office to maintain a critical approach when assessing the relationship between copyright and machine learning and to base any legal analyses on the factual technological underpinnings of how works are used to develop artificial intelligence systems.

To that end, CDT and the R Street Institute offer the expertise of the GRAIL Network (Governance Research in Artificial Intelligence Leadership Network), a group of computer scientists, researchers, economists, and legal experts working on a broad set of issues related to the development and use of AI.[1] The network intends to bring policymakers and technical experts together to better align policy development with technical reality and to help technical experts play a more active role in policymaking. We encourage the Office to consider GRAIL as a resource for a better understanding of the technical, legal, and societal implications of AI.

We address the USPTO's questions relating to datasets and databases below.

3. To the extent an AI algorithm or process learns its function(s) by ingesting large volumes of copyrighted material, does the existing statutory language (e.g., the fair use doctrine) and related case law adequately address the legality of making such use? Should authors be recognized for this type of use of their works? If so, how?

[1] See www.grailnetwork.org or contact CDT for more information.

Fair use and related case law adequately address the legality of using copyrighted material to train machine learning systems.

As a primary matter, CDT suggests that answering this question begins with a careful look at what the "use" of copyrighted material means in the context of machine learning. When machine learning systems "ingest" data, they are measuring various aspects of the individual elements and mapping the relationships among them. More simply, they extract facts about the data and look for patterns among those facts. It is through this process that systems discern patterns, correlations, and logical relationships in the dataset, and it is from these relationships that rules and predictive models are derived and refined. The use, then, of a copyrighted work is as a source of information representing one or more data points in a larger collection of data points. This kind of use is fundamentally different than a use where some or all of an original work is reproduced as part of the new work.[2]

Based on this understanding of the use of data in machine learning, CDT suggests that where data elements are protected by copyright, and those elements are reproduced at some stage of the process, such use would be best classified as either "research" or "teaching," both of which are explicitly listed as examples of fair use in Section 107.[3] The act of measuring and extracting facts about data elements, if performed by humans, would certainly fall into the category of research because it is the methodical search for new information *about* the data in question. Likewise, reproducing copyrighted data elements for the purpose of developing, refining, or testing an algorithmic model could equally be construed as "teaching." For example, in some methods of machine learning, programs are instructed as to the relationships between data elements and related information. One familiar example of supervised learning is the "mark all the boxes containing x" exercise used by some websites to deter bot users, in which a human identifies and tags portions of images with identifying labels.[4]

Given the wide variety of fact patterns for using copyrighted data elements to train or test automated systems, a single fair use analysis according to the four-part statutory test may not accurately represent all possible outcomes. However, we offer the following observations about some aspects many such uses will have in common, with respect to the fair use test.

---

[2] CDT notes that authors and judges have discussed other computer-related uses of copyrighted works as "non-expressive," and would agree that the use in question here is also non-expressive. See, e.g. James Grimmelmann, *Copyright for Literate Robots*, Iowa Law Review, Vol. 101:657 (2016); *Authors Guild v. Google, Inc*. 804 F.3d 202 (2nd. Cir. 2015). However, using copyrighted works in machine learning differs from those uses because most algorithmic models retain none of the content of the works from which they are derived.

[3] 17 U.S.C. §107.

[4] *See*, e.g., James O'Malley, *Captcha if you can: How you've been training AI for years without realizing it*, TechRadar (Jan. 12, 2018) https://www.techradar.com/news/captcha-if-you-can-how-youve-been-training-ai-for-years-without-realising-it

Factor 1: Purpose and character of use- As noted above, use of copyrighted data elements for developing or training a machine learning system is always for the purpose of either research or teaching. This should weigh strongly in favor of a finding of fair use. CDT is unaware of any cases in which a developer of a machine learning system markets any reproductions or portions of copyrighted data elements used in development. However, to the extent that the ultimate application of the model developed using the copyrighted data elements is commercial in nature, CDT suggests that such commercial use would only weigh against a finding of fair use when the application negatively impacts the market for the work (under Factor 4). We also note the several degrees of separation between the use of a copyrighted work as one data element in a training corpus and the outputs of an automated system in which a model derived from the training data is used. As this relationship between the work used and the ultimate output of the model becomes less direct, the weight assigned to any commercial uses should decrease.

Factor 2: Nature of copyrighted work- Given the diversity of copyrighted works that may be used as a corpus for machine learning, general analysis under this factor is difficult. However, we note that even for the most creative works of artistic expression, it is the information derived *about* the work and its relationship to other data elements that is relevant to building, training, or testing a model. This analytical approach attenuates the importance of a work's creativity. From this perspective, the nature of the copyrighted work should not weigh strongly in the analysis.

Factor 3: Amount and substantiality of portion used- As with the second factor, a uniform assessment under Factor 3 is not possible given the diversity of training methods used for machine learning. But even assuming that entire copyrighted works are reproduced in the training process, this factor would not necessarily weigh against a finding of fair use.[5] Moreover, in most cases no part of any reproductions made in the training or testing of a model become part of the final product. Rather, the models or rules derived from the training data reflect the relationships or commonalities found among the data elements. From this perspective, the works used are a source of information but the content of the works is not "borrowed" in the way that a traditional fair use assessment contemplates. Instead, the use of copyrighted works in machine learning is more like an art student studying the paintings of great masters in a museum and incorporating their impressions of those works into their mental model of what classical paintings look like. Although reproductions may be made in the process, either in the mind of the student or the memory of the computer, the content of the reproduced work does not become part of the final product. Unless some or all of the original work remains in the system, this factor should not weigh against a finding of fair use, regardless of the portion used in the training or testing of a model.

Factor 4: Impact on the market for the work- The wide variety of machine learning applications again makes a uniform fair use assessment difficult. CDT is aware of some generative

[5] *Authors Guild v. Google, Inc*. 804 F.3d at 221 (2nd. Cir. 2015) (Noting that although Google copied entire works, it did not make them available to the public in a way that undermined the authors' interests in distribution.)

adversarial networks (GANs) that are capable of producing visual and literary works.[6] Although it is unclear whether those works have any effect on the market for or value of the works used to train the network, we note that the relationship between the use of copyrighted works to train a GAN and the market for either the original works or the GAN-produced works is indirect at best. As with the other factors, the "use" here is not of the same kind that traditional fair use assessments contemplate, where portions of one work are re-used in another work. In most cases, exposing a machine learning system to works does not result in portions of the training dataset becoming embedded in either the models or any output of the system.[7] Instead, it is information—facts—about the training dataset that shape the models and the resulting outputs.

Fair use offers flexibility.

Despite the unconventional use case machine learning presents, CDT believes fair use adequately addresses the legality of such uses. Further, fair use offers advantages over other possible legal mechanisms for allowing the use of copyrighted works in the context of machine learning. For example, in places where databases and datasets have additional use restrictions, such as the *sui generis* database protection in the European Union, policymakers have been pressed to create exceptions to accommodate machine learning.[8] Yet despite prolonged negotiations, the resulting exception for text and data mining (TDM) is so rigid and restrictive as to prevent many beneficial uses of datasets.[9]

In particular, the TDM exceptions suffer from two major flaws. First, the exception created in Article 3 limits the uses to "non-commercial" uses by a narrow set of "research organisations and cultural heritage institutions."[10] These limitations prevent many legitimate uses of copyrighted works in datasets that pose no risk to authors' ability to exploit their own works. Second, Article 4, which is more permissive in terms of uses and users, allows authors to deny use of their works.[11] This ability creates logistical and practical barriers to uses of large datasets in which some, but not all, authors have opted out of the exception because the additional efforts required to identify, cull, or negotiate individual licences for works make using the dataset

---

[6] *See*, e.g. *These Works of Art Were Produced By Artificial Intelligence*, Duke Today, (Mar. 18, 2019) https://today.duke.edu/2019/03/these-works-art-were-created-artificial-intelligence.

[7] CDT acknowledges that some outputs of GAN systems may be similar to existing copyrighted works, but we take no position as to the legal implications of that possibility as a general matter.

[8] Directive (EU) 2019/790 of the European Parliament and of the Council on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC, Art. 3, 4 (Apr. 17, 2019). ("DSM Directive")

[9] *See*, CDT, CDT's Concerns on the European Commission's Proposal for a Directive on Copyright in the Digital Single Market, (2017), https://cdt.org/wp-content/uploads/2017/02/CDT_Concerns_EC_Proposal_Directive_on_Copyright_DSM.pdf. *See also*, 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/

[10] DSM Directive, Art. 3(1).

[11] DSM Directive, Art. 4(3).

unfeasible. In comparison, fair use allows more legitimate, non-infringing uses of copyrighted works while still protecting the interests of the rightsholders and does not suffer from the rigid definitional limitations or the practical barriers presented by statutory exceptions such as Art. 3 and 4.[12]

Authors need no additional recognition for this type of use of their works.

Given the non-expressive, non-exploitive nature of using copyrighted works in the development of algorithmic models and other machine learning tools, it is unclear why or what kind of additional recognition authors need when their works are included in data sets used for machine learning. In CDT's experience, the most commonly used datasets involving copyrighted works are composed of works for which authors have granted some type of general license.[13] For many of these licenses, no additional recognition is required. Others, such as the CC-BY license, require attribution when works are reproduced, but such attribution makes little sense when reproductions are not made publicly available, but are instead made for the convenience of the researchers and developers working with the dataset. To require public attribution for each author of a work in a dataset from which a model was derived would be like asking a novelist to list each and every author whose works they read at any time before drafting, for each would represent a datapoint in the author's mental model of the appropriate content and structure of a novel. CDT sees no reason to pursue this result in the machine learning context.

9. How, if at all, does AI impact the need to protect databases and data sets? Are existing laws adequate to protect such data?

There is no need for further intellectual property protection for databases or datasets.

Individual works within data sets are already protected against unauthorized, exploitative uses. Those protections address the incentives a human-centric IP system intends—motivating the creation of new works by protecting authors' ability to monetize them. Existing copyright law already inhibits the use of works for machine learning; adding additional protections for particular arrangements of the data, on top of the existing rights for the authors of individual copyrighted elements, would further hinder, rather than promote, the progress of science and the useful arts.[14]

First, to the extent that researchers and developers of automated systems find value in particular arrangements or curations of databases or data sets, those arrangements and

---

[12] *See*, Vadym Kublik, *EU/US Copyright Law and Implications on ML Training Data*, Valohai, https://blog.valohai.com/copyright-laws-and-machine-learning, last visited Jan. 10, 2020.
[13] See, *Common License Types for Datasets*, https://help.data.world/hc/en-us/articles/115006114287-Common-license-types-for-datasets, last visited Jan. 10, 2020.
[14] *See* Arjan Wijnveen, *How copyright is causing a decay in public data sets*, LinkedIn, Nov. 28, 2916, https://www.linkedin.com/pulse/how-copyright-causing-decay-public-datasets-arjan-wijnveen.

curation efforts are best made by the researchers and developers to suit their particular needs. But the first party to arrange or curate a data compilation should not be able to prohibit others from rearranging or differently curating that database, nor should they be able to prevent others from arranging the same data in the same fashion. As discussed above, jurisdictions with additional protection for databases are now facing difficulty reconciling their *sui generis* rights with the need to access and use data. CDT strongly suggests that copyright policy in the United States should avoid this problem by not extending additional protections to databases.

Second, further restricting access to and use of databases and data sets will limit researchers' ability to reduce bias in algorithmic models. The lack of access to sufficiently representative data is a common contributor to biased algorithms.[15] Conversely, access to more diverse data sets helps researchers test for and mitigate bias in automated systems.[16] Even under existing copyright law, the "friction" associated with using some of the largest, most comprehensive data sets steers researchers and developers toward data sets that are easier to access and involve less legal risk.[17] This also shrinks the world of data sets available to researchers, which is a bias in of itself, regardless of the relative bias expressed in the datasets themselves. Additional protections for data sets would only increase that friction, making it even more difficult to combat bias in algorithmic models.

Finally, although there are other concerns with access to and use of data sets, such as preserving the privacy and confidentiality of certain data, those are outside the scope of copyright. CDT acknowledges that certain use restrictions may be appropriate for some kinds of data to ensure that the rights of data subjects are not infringed. However, those use restrictions need not be based in copyright. In fact, copyright is a poor tool to protect the rights of people whose personal information may be reflected in data sets because in many cases, the data subject and the copyright holder (if any) are not the same entity. Although we disagree that additional copyright protections are warranted for data sets (beyond any protections granted to the works therein), we strongly encourage the Office limit any proposed expansions of intellectual property protections to only those which directly advance the purpose of copyright.

Respectfully submitted,

Stan Adams
Deputy General Counsel & Open Internet Counsel
Center for Democracy & Technology
1401 K St. NW, Suite 200
Washington, DC 20005

[15] See, generally, Amanda Levendowski, *How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem*, Washington Law Review, Vol. 93: 579 (2018).
[16] *Id.*
[17] *Id.* at 593.