Using Fair Use to Reduce Algorithmic Bias

Author : Michael W. Carroll

Date : June 28, 2019

Amanda Levendowski, How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem, 93 **Wash. L. Rev.** 579 (2018).

What is the relationship between copyright law and artificial intelligence or machine learning systems that produce outputs biased by race, gender, national origin, and related aspects of being human? That is the question that Amanda Levendowski investigates and addresses in her refreshingly well-written, to-the-point article *How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem*. In a nutshell, she argues that: (1) these systems need large quantities of training data to be effective; (2) those building these systems rely on biased data in part because of their own biases but also because of potential risks of copyright infringement; and (3) more copyrighted works can legally be included as training data under the fair use doctrine and should be so used to selectively diversify the inputs to these systems to de-bias their outputs.

Levendowski starts with the problem in the form of Google's natural language processing system word2vec. It is a form of neural word embedding that analyzes the context in which words appear in the source texts to produce "vectors," which indicate word associations such as "Beijing" is to "China" as "Warsaw" is to "Poland." Trained by analyzing the published news sources incorporated into Google News to which Google has obtained a copyright license, word2vec ingests the biases in those sources and spits out results like "man" is to "computer programmer" as "woman" is to "homemaker." Levendowski acknowledges that those in the machine learning research community agree that this is a problem and are in search of a solution (including Google's own researchers)[1], but she responds that it should not be left only to developers at large technology companies with access to the training data to de-bias their own systems.

Levendowski further asserts that copyright law stands as a potential barrier, or at least a perceived barrier, to outside researchers' ability to investigate and report on bias in these systems. Copyright reinforces incumbents' advantages in three ways. First, while reverse engineering of the algorithms is protected by fair use, accessing those algorithms, if they are subject to technological protection measures under 17 U.S.C. §1201, is limited to the narrower § 1201(f) exception or the right to circumvent that the First Amendment may provide.[2] Second, if a biased system's underlying training data is copyrighted, journalists and other investigators who seek to expose the sources of algorithmic bias are likely to be chilled by the prospect of an infringement suit. Finally, the leading artificial intelligence developers have significant resource advantages that allow them to acquire enormous training datasets by building them (Facebook) or buying them (IBM).

This competitive advantage leads newcomers to rely on what Levendowski terms "biased, low-friction data" or BFLD; that is, data that are accessible and that carry little legal risk. (P. 589.) Here, her example is the 1.6 million emails among Enron employees made accessible by the Federal Energy Regulatory Commission in 2003. This is one of the only publicly-accessible large datasets of interlinked emails. Although these emails are technically works of authorship protected by copyright, the legal risk that any of these authors would sue an AI researcher for using these is close to nil. But, this is hardly a representative sample of people to study if one were to train a system to extract generalizable rules about how human beings communicate by email. Other examples of BFLD that have other forms of bias include public domain works published prior to 1923, which do not reflect modern language usage, and Wikipedia, which is legally low-risk because of its Creative Commons license but is a biased source of facts about the world because of the large gender imbalance among contributors. Levendowski argues that this imbalance biases the data in the language used to describe women in many Wikipedia entries, and the substance of these reflect male bias in terms of the subject

matter covered and the subject matter omitted, such as key facts about women in biographical entries.

The article then argues that enlarging any of these datasets, specifically with diverse, copyrighted sources that are likely to mitigate or erase bias, is desirable and is legal as a fair use. Recognizing that access to these sources remains a challenge, Levendowski argues that at least the use of these sources should be cleared by fair use.

Here, I should disclose my bias. I have a forthcoming article that makes a related argument that copyright law permits the use of large sets of copyrighted works for text and data mining, so I am sympathetic to this article's argument. Nonetheless, I think most readers will find that although the fair use analysis in this article is brief, perhaps too brief, it is supported by the case law and copyright policy.

The analysis argues that using copyrighted works as training data is a transformative use, and there is now substantial case law and scholarship that support this assertion. The use is for a different purpose than for which the works were published and the use adds something new through the system's operation. The article then argues the second factor also favors the use because even creative works are being used for their "factual" nature; i.e., as examples of creative works by humans. Under the third factor, using the entirety of these works is necessary and appropriate for this purpose and has been approved in a number of cases involving computational processing of copyrighted works. Finally, under the fourth factor, even if some of the training data has been licensed in by current developers, the transformative purpose under the first factor overrides any negative impact that fair use may have on this market.

While this analysis is generally persuasive, I found this part of the article a little thin. I agree that a court would almost certainly characterize this use as transformative for the reasons stated. But, the second factor has traditionally been focused on how much expressive material is in the work being borrowed from rather than the borrower's purpose. This move felt like giving the transformative purpose a second bite at the apple. While the second fair use factor does little work on its own, I think it is appropriate to consider as part of the balance how much original expression is at stake.

I will note that I wanted more discussion of the third and fourth factors. While it is easy to agree that use of entire works is likely to be permissible, the harder question is how much of that training data can be made publicly available under fair use by those seeking algorithmic accountability. I would have liked to know more about how and where Levendowski would draw this line. Similarly, the evidence of some licensing for this use, needs more elaborate discussion. I agree that the transformative purpose is likely to insulate this use, and that this licensing market is really one for access to, rather than use of, the training data, which diminishes the impact under the fourth factor.[3]

With that said, I want to acknowledge the creativity of Levedowski's thesis, and show appreciation for her clear, succinct presentation of the three stages of her analysis. This piece is a welcome contribution by an early-career researcher, and I look forward to reading her future work.

1. Ben Packer et al., Text Embedding Models Contain Bias. Here's Why That Matters, **Google Developers Blog** (Apr. 13, 2018).
2. *See* Universal City Studios, Inc. v. Corley, 273 F.3d 429 (2d Cir. 2001)(recognizing that §1201 can be applied in a way that burdens speech and is subject to intermediate scrutiny when it does so).
3. Here I want to recognize the argument advanced by Ben Sobel on the fourth fair use factor. He argues that, at least when creative works are used to train systems designed to create competing creative works, the fourth fair use factor should weigh against such use. *See* Benjamin L.W. Sobel, Artificial Intelligence's Fair Use Crisis, 41 **Colum. J.L. & Arts** 45, 75-79 (2017). It is a creative argument by which I am not persuaded because fair use should not be the policy finger in the dike holding back automated creation in my view. But, I found his arguments about the ways in which machine learning systems may require more nuanced fair use analysis to be well made.

Cite as: Michael W. Carroll, *Using Fair Use to Reduce Algorithmic Bias*, JOTWELL (June 28, 2019) (reviewing Amanda

Levendowski, *How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem*, 93 **Wash. L. Rev.** 579 (2018)), https://ip.jotwell.com/using-fair-use-to-reduce-algorithmic-bias/.