30 October 2023

I am a social scientist with expertise in digital technologies and in language and technology, as well as an award-winning writer of speculative fiction and poetry. I write now both as a scholar and an author. I focus these comments primarily on the copyright harms of "AI" tools with regard to fiction and the publishing industry, though the underlying points apply more broadly. These comments are necessarily incomplete and not intended to be exhaustive.

I urge you to uphold the rights of individual copyright holders—those whose fiction, essays, and poetry technology companies have already used unlawfully as training data and input data to derive their language models (LMs); and those whose works will be vulnerable to future use as companies making transformer-based LMs such as ChatGPT seek ever larger datasets. Remedies are needed for the former, robust protection for the latter.

Starting Points

As one starting point, companies and other entities building such models should provide documentation of the specific data used, both with regard to training data and input data, that can be easily audited to ensure that they are not violating copyright. Some companies will likely complain that such documentation is too large an undertaking or one too late to accomplish now. Perhaps such protests might have been persuasive in the early 2000s, but it is 2023 and we have been having extensive, nuanced conversations about data ethics, data protection, data sharing, and the data spectrum for many years now. This is indeed a large undertaking, and it is one that companies should have been doing from the beginning. That they chose not to speaks to their belief they can exploit and extract without consequence. It does not mean that we should accept their delusion. There is a clear and well-established procedure for licensing the rights of published works; technology companies chose to disregard it.

Which brings me to another point: At the moment, some companies offer authors an option to opt their works out of an input dataset. This is insufficient. The administrative burden for removal must lie on those who created the datasets and chose to ignore copyrights. Technology companies should remove all works that are not in public domain and for which they do not have explicit consent (actual consent, not just assent tucked into a clickwrap terms of service).

Further, based on what is currently known about the data used, copyrighted materials have not only been used in input datasets, they have likely also been used in *training* datasets. (LMs are "pretrained" before use with an input dataset.) That means that even those LMs that use input datasets that consist solely of public domain materials may still be using copyrighted materials in their training data. Further, even with the removal of all copyrighted materials from both the input data and training data, the development of LMs, particularly transformer-based LMs such as ChatGPT (often known as LLMs), has already benefitted considerably from extensive use of copyrighted materials. Remedies will need to reckon with this gain from unlawful use as well.

Fair Use

Some companies will undoubtedly argue that their use of copyrighted material constitutes fair use. I believe this should be rejected for multiple reasons.

First, fair use is a defense of the use of *selections* from a work—in these cases, technology companies appear to have used entire works. Companies will likely argue that the quantity of the input—that consideration of the input—is less important than the nature of the output. (I will return to the output.) But consideration of the input has to be a primary concern. As a writer, I pay for the materials that I use, whether that's paper, the electricity that powers my laptop, or access to research. If companies are using more than very limited selections from copyrighted materials to build language models, they need to pay for them, just as any other creator pays for materials. Even if they were to use only very limited selections, there's a reasonable question whether that would constitute fair use, whether these technologies reuse copyrighted materials to comment or critique, or for educational purposes.

I will not delve here into the question of whether using an LM-based chatbot to access specific details of copyrighted material constitutes infringement, as that is the subject of ongoing litigation and I strongly suspect people with more knowledge of those aspects than I will be submitting comments. But I will underscore that, though such chatbot regurgitation has commanded much of the headlines about these technologies and copyright, producing, for example, the nickname "plagiarism machines," that is a different—though obviously related—element of use than involved in training or inputting data into a language model.

Often, LMs are described as extracting language relationships from data, as if language relationships were somehow an inherent property of written texts, divorced from the humans who wrote those texts. The art of authors is quite literally crafting language relationships. This is one of the reasons these companies deem such data "high quality" and prefer it, for example, to social media posts. Indeed, if this compositional artistry were not one of the key aspects that LMs were extracting, companies could successfully train models just on grammar textbooks and the dictionary.

A company booster would likely argue that a single work is insufficient to build a language model, that copyrighted materials must be used in aggregate to discern robust relationships—and that this equates to no single work's artistry being meaningful to the process. I suggest, however, that this is the wrong framing. Each work is indeed meaningful, for its meaning is not measured by a technological process or the current limitations of such a process. The value and meaning of the work lie in the work itself. Companies may wish to combine a work with other works to explore patterns across them—in fiction publishing we usually call this an anthology—if so, companies still need to pay copyright holders to do so.

Another reason I do not believe companies' use of copyright material qualifies as fair use is the commercial effects. Here, I shift focus from inputs to outputs. Fair use has always been a context-sensitive doctrine; I suggest that, in the case of LMs, while inputs should be considered individually—for they are individual acts of artistry and authorship—when we consider the

technologies and their outputs, we must consider their effects at a broader industry level, for these technologies are, after all, specifically designed to scale.

And when we do, we see: 1) their outputs directly compete with short fiction and novels (an assortment of novels generated using LMs are already for sale on Amazon); 2) the technology is designed to deskill the writing profession; making a living as a professional fiction author is already extraordinarily difficult and if publishers can embrace these outputs, it will reshape labor relationships to the detriment of authors and the advantage of large companies; novelists, of course, cannot presently unionize in the US, but such technologies' deskilling effects constituted one of the specific concerns that prompted the WAG to strike; and 3) these technologies work to shift value from an author's work to a company's paratexts (see, e.g., YouAI's "conversational companions" designed to accompany a book and regurgitate its contents via chatbot), much in the way that Google Search's preview feature has shifted web traffic away from the websites providing the information to Google itself. Across the industry, these effects harm the commercial viability and value of novels written by human writers without the use of LMs.

Perhaps most importantly, these technologies not only do *not* support the progress of the arts—the fundamental aim of copyright—they are by nature regressive. Though technology companies claim them transformative, they are transformative only in their attempts to transform labor relations. As Bender and Gebru et al (2021) note, LMs are merely stochastic parrots, assemblages of language predictions based on past texts. They are always, inherently, facing backwards, pointed toward historical data. This undermines our society's trajectories of artistic exploration. And there is no need for such technologies: fiction publishing does not lack for writers and writers do not lack for ideas.

Further, as many studies have already detailed, these technologies resurface and reinforce biases. Fiction and publishing in the US have a long history of racism, misogyny, homophobia, etc. Indeed, despite gains in recent years, struggling against such forms of supremacy continues very much today. Having to refight, for example, the idea that rape is an acceptable step in building a relationship, as was prevalent in romance novels until the 1990s or so, is not progress. To say nothing of the harms a fresh wave of older slurs and stereotyped characters would yield. And yet, these are very real concerns if these technologies were to become pervasive in fiction publishing.

For all of these reasons, I urge you to uphold the rights of individual copyright holders, with appropriate remedies and protections, and not to reframe companies' theft as fair use.