Contents

1 'Original Expression' to 'Aggregate Forms': Language and Copyright in the Shadow of AI

1.1 draft

1.1.1 introduction

I'm going to speak on this topic of open data "in the shadow of AI" from the perspective of someone who actively uses ML tools in her research and her teaching.

As a programmer, who was trained in Literary Studies and DH, I am a strong proponent of trying to find sustainable and ethical ways of working with this technology. For me, that means defining "open" in a way that supports the sharing of knowledge resources to empower not just the "commons," but those who most need it.

For many years now, as many people know, AI companies are scraping as much of the web as possible in order to train their ever larger language models. However, what many people don't know, or don't hear much about, is that the question of which data is legally scrapable is being debated in the courts. Right now, there are over 40 ongoing lawsuits brought by content creators against companies like OpenAI, Anthropic, Midjourney, and many others.

SLIDE court cases

Here you can see a list of some of these cases, the ones that are just in the city of New York, which is where I'm based.

To defend themselves, these companies argue that their data harvesting constitutes "fair use," which is a copyright protection that allows the taking of original work depending the purpose and nature of the use, and the effect on the market for the original work, among factors.

For example, in the lawsuit brought by the NYTimes against OpenAI, OpenAI defends itself by saying they are using copyrighted data to "learn

facts" about language, which is a purpose that is protected under "fair use."

They claim that,

SLIDE quote in memorandum of law

> ... it is fair use under copyright law to use publicly accessible content to train generative AI models to learn about language, grammar, and syntax, and to understand the facts that constitute humans' collective knowledge. OpenAI and the other defendants in these lawsuits will ultimately prevail because no one—-not even the New York Times—-gets to monopolize facts or the rules of language.
>
> - MEMORANDUM OF LAW filed by OpenAI, February 2024

So, for this presentation, I'm going to focus this phrase, the "facts and rules of language", and the claim that this is all that language models extract from data during the training process.

I will assess this claim by examining what happens to language data when it is ingested by a ML algorithm. To do so, I will offer examples from my own research, where I develop language models for the purpose of studying social bias, specifically anti-trans bias, which is a major issue right now in the United States.

By exploring what happens to language data within ML systems, I hope to create some context for considering how we might protect digital language forms from mass extraction, while simultaneously making them open to certain uses.

1.1.2 the defense of "transformative"

First, I'm going to examine OpenAI's argumentation more closely, to understand what they mean by the "facts or rules" of language.

Their argumentation begins by appealing to a stipulation from the Fair Use clause, which emphasizes the importance of "transformative-ness", that is, how much the derivative object changes from the original. They cite a legal case from 2014, *Authors Guild v. HathiTrust*, which is about the legality of search results. The ruling for this case finds that search results are fundamentally transformative from the works that they reference, because they offer a new object, that is, *information about works* in their database. This ruling determines that objects that represent an *aggregation*, or language as an *aggregate form*, is significantly different from the original as to constitute Fair Use.

OpenAI applies this understanding of "transformative" to their language models. They argue that, like search results, language models constitute a new kind of object, one that *generalize language patterns*.

SLIDE OpenAI quote

They explain that,

> "By learning patterns from its training corpus, an AI system can eventually generate media that shares some commonalities with works in the corpus (in the same way that English sentences share some commonalities with each other by sharing a common grammar and vocabulary) but cannot be found in it." ("Comment", 9-10)

(I will leave aside for the moment, that while grammar and vocabulary are not copyrighted, dictionaries and grammar books are indeed copyrighted material.)

But moving on, what strikes me is this use of this word, "patterns," and what that could mean in practice. It seems that they are referring to some ideal of language, of language as a formula or structure, which is distinct from the specifics, the content, of language.

Now, I want to ask, how does this characterization of language weigh against these models' operatation in practice, and their outputs?

1.1.3 close reading

To explore this question, I'm going to show a few examples from my current research, which uses ML tools for the purpose of studying discriminatory and prejudiced language. For this, I scrape data from various internet sources, which represent different perspectives. Then, I use this data to train (very small) ML models, to see how they respond to certain prompts.

The goal here is not to generate text for the sake of generating text, but to surface certain insights about the data on which that text was trained.

I'm going to show some examples of ML-generated text based on two very different data sources, representing different political perspectives on the topic of transgender rights in the US.

SLIDE heritage screenshot

One of these sources represents an American conservative perspective, and comes from the Heritage Foundation, which is a think-tank in Washington DC whose goal is to influence governmental policy. You can see some of the headlines here from their website, like "Sorry Democrats, but Trumps'

'Two Sexes' Executive Order is Constitutional". All of these articles, about 300 in total, were included in my dataset.

SLIDE ACLU trans screenshot

Coming from the opposing side, from the progressive pole, is the ACLU, the American Civil Liberties Union, which is a group of legal professionals and volunteers who advocate on behalf of civil rights for marginalized groups in the US. Here, you'll see articles that speak of trans in terms of "rights" and "liberation," and connect it to LGB rights more broadly.

With these datasets, I then trained two individual large language models, using gpt-2 (an open source model) as the base model. I'll mention quickly, for those who don't know, that training models happens in various stages. What I did was take an already trained model, gpt2, and re-trained it on a smaller and more specific dataset. This is technically called "fine-tuning", and its much easier and less resource intensive than training the underlying "base" model (I am more than happy to explain more details in the Q&A).

So, after training, I fed a series of prompts to both of the resulting models.

These prompts included:

SLIDE prompts

Masculinity is

Femininity is

Transgender is

Gender binary is

Man is

Woman is

And finally, I comapared the results.

First, there was a strong contrast of gender and how genders are conceptualized between the model trained on the ACLU data and model trained on the Heritage data. First, I'll show some examples from the ACLU model, which represents the progressive side:

Masculinity is a matter of love and celebration.

Masculinity is a space for hope and liberation for all.

Masculinity is not defined solely by the beauty of our bodies, but by the beauty of our experiences.

Femininity is a celebration of beauty, feminine liberation, and femininity.

> Femininity is our joy, our struggle, and our fight is our struggle.
>
> Femininity is about allowing people to express themselves without government interference.

As you can see, terms associated with "masculinity" and "femininity" are characterized by gender-affirming, even celebratory language, which is very positive and empowering; using words like "liberation," "beauty", and "joy". Reading this, you get the sense that these terms come from contexts mentioning transgender experiences, and in the training data, "masculinity" and "femininity" may have been prepended by the term "trans."

From these examples, you may also notice the tendency of language models (especially very small and underdeveloped ones) to repeat themselves. This is a quirk due to their predictive nature, where the goal is to guess the next word based on what is most likely. As a result, they get themselves stuck into these little loops of saying the same thing over and over again—which I find kind of endearing.

Now, I'll show the text generated by the model trained on the Heritage Foundation data.

> Masculinity is the cornerstone of Western civilization.
>
> Masculinity is the fruit of patriarchy, and patriarchy is the heart of conservatism.
>
> Masculinity is defined by the ability to produce sperm, eggs, and live children.
>
> Femininity is an enduring American tradition.
>
> Femininity is defined by means of the relationship between the sexes, the ability to raise their children, the capacity to provide for their own reproduction, the capacity to provide for their own children, the ability to provide for their own.

Here, these gender terms are also positive, but their associations with culture, tradition, and reproduction—things that suggest stability rather than empowerment.

So you can see, even from just glancing at the results, that there are direct connections between the training data and the model outputs. By processing the training data, the model learned not just how language works, the "facts or rules" of language, but absorbed the perspectives contained within that language. So that, depending on the dataset that the model is trained on, the terms "masculinity" and "femininity" will have totally different meanings.

Perhaps this is not surprising. But what I also found, which deepens this a little bit, is that gendered terms reveal investments in other, seemingly unrelated or benign terms.

SLIDE subjective

For example, the Heritage Foundation model kept repeating the term "subjectivity" when it mentions gender:

> Masculinity is a subjective self-perception, not a universal concept.
>
> Femininity is a subjective, internal sense of self.
>
> The gender binary is a subjective, malleable, and often incorrect idea.
>
> The gender binary is a subjective, internal, and often transitory concept.
>
> The gender binary is a subjective, grammatically incorrect and illogical concept that conflates sex and gender identity.

If you're familiar with American conservative viewpoint—that gender binary is based firmly on biology—you may notice that these examples don't reflect that view. Rather, they represent the a progressive view of the gender binary which asserts that gender is based on aspects of identity beyond just biology.

The reason for this, I believe, is that this particular term, "subjective" *does not* describe the conservative position. Rather, it describes a conservative frame for the progressive position. In other words, it represents what a transphobic person thinks a progressive person thinks gender is—as something insubstantial, as a feeling.

This explains why there is a curious hint of contempt in some of the examples, which use terms like "illogical" and "incorrect" alongside "subjective." These are traces of derision which are sustained from the training data.

In the model outputs then, we see not just a single perspective of gender, but a *flattening* of perspectives into a single statement. From the training data, there are distinct expressions, distinct viewpoints, which in the model, have been aggregated into an apparently univocal utterance.

Clearly, this language presents a different kind of object from that which OpenAI claims falls under the protection of "fair use." In addition to absorbing the "rules and facts" of language, the model also takes up the perspective of the training data. And, crucially, depending on whose viewpoints have

been absorbed into the language model, this can be dangerous for vulnerable groups, like trans people.

So, language in this form, in an *aggregate form*, the so called "facts and rules" of language come hand in hand with specific perspectives, and sometimes distinct perspectives.

This flattening or amalgamation of perspectives has major effect on how we might approach data in the "shadow of AI," taking a quote from this panel's title.

While existing copyright law seeks to protect expression, what is sometimes called "original expression" and sometimes called "intellectual property," maybe, in the shadow of AI, we ought to think more about protecting certain groups or communities which the data pertains to.

There is some compelling work exploring this area, especially coming from data sovereignty movements in the global south. These groups offer new data licensing schemes that consider data rights in terms of the community. They offer some inspiration for thinking about how we might move toward more need-based forms of data licensing.

SLIDE nwulite license

For example, the Nwulite Obodo license, which was developed by a team of researchers in Pretoria, South Africa, offers different tiers of permissions based on who is using the data. Users who are from developing countries can use the model freely, while other users must either pay or commit to releasing their derivatives under the same license.

SLIDE questions

They share some questions which helps to identify the level of access that potential users ought to have:

> How do I ensure that others like me who want to use outputs and other datasets from my project/work are able to do so?
>
> Should those who own or have effective and working access to compute and other infrastructure have the same kind of access to my outputs as those who do not have?

Here, the idea is prioritizing consideration of the resources and purposes of those who want to use the data. So that the permissions for big tech company, would differ from those to a community educator.

Licenses like the Nwulite Odobo are specific to the vulnerabilities of the groups they pertain to (in this case, to speakers of African languages), and I don't think that would work for any case, such as for data about trans people.

But the way they think about data, prioritizing questions about access and community, over originality or "expression," makes a lot of sense for language that takes aggregate forms.

Thank you.