Contents

1 draft

1.1 introduction

I'm going to speak on this topic of open data from my perspective of a programmer, as a person who actively uses ML tools in her research and her teaching.

While these tools are deeply problematic, for many reasons, and most urgently, environmental reasons, I am a strong proponent of trying to find sustainable and ethical ways (if there can be any) of working with this technology.

And when it comes to digitized data, and ecologies of open and openly accessible data, that means defining "open" in a way to sustain the sharing of knowledge resources that empowers society as a whole.

For AI companies, anything that is accessible on a webpage is considered open to extraction and copying, for the purpose of training their ever-larger language models toward the goal of profit.

And right now, in the United States, the question of which data can be legally harvested by AI companies to use for training their models is being debated in the courts. ~~There are over 40 ongoing lawsuits brought by content creators against companies like OpenAI, Anthropic, Midjourney, and many others, which allege that these companies have committed copyright infringement by taking openly accessible data and privatizing the resulting products.~~

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.

SLIDE The New York Times Company v. Microsoft Corporation, December 27th, 2023

For example, in the lawsuit brought by the NYTimes against Microsoft (the owner of OpenAI), OpenAI defends itself by saying that data harvesting

is for the purpose of "learning 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 assess this argument from a technical perspective, by examining excatly what happens when to language data when it is ingested by a language model in order to generate output. I'm going to offer examples from my own research, where I develop ML models for the purpose of studying social bias in language, specifically anti-trans discourse, which is a powerful movement right now in the United States.

I'm going to spend some time going over the technical details of ML processes, because, in my view, tech companies gain too much power from the widespread ignorance about how ML tools actually work.

So, by the end I hope I will have left you with some context of what happens to training data within ML systems, and what kind of language forms it creates. Then, I hope in the Q&A, I hope we can talk about potential ways for protecting digital language forms from mass extraction and data colonialism.

1.2 the defense of "transformative"

First, I'm going to examine OpenAI's argumentation more closely, to trace what exactly 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 *transforms* from the original object. They cite a legal case from 2014, *Authors Guild v. HathiTrust*, whose ruling finds that the database format of search results is fundamentally transformative from the works contained within the database, because it offers a new object, that is, *information about works* in the database. This ruling determines that

search engines are permissable, by determining that language as *aggregation*, or an *aggregate form*, is fundamentally distinct from langauge in its original, syntactic context.

OpenAI applies this understanding of "transformative" to their language models. They argue that, like search results, language models constitute a new kind of object, which *generalizes language patterns* from the original.

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 is 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?

1.3 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, using ML as a kind of text analysis tool. For this, I gather and create custom datasets from various internet sources, which represent different perspectives. Then, I use these datasets to train (very rudimentary and 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.

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. From their website, I scraped all the articles that were organized under the topic heading of "gender", of which you can see some of the headlines here. You can tell the kind of perspective and tone by some of these headlines, like the one that says, "Sorry Democrats, but Trumps' 'Two Sexes' Executive Order is Constitutional".

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 that training models happens in various stages, so what I did was take an already trained model, gpt2, and re-trained it on a smaller and more specific dataset. This is a process that's technically called "fine-tuning" (And if people are curious about the specifics about this process, I am more than happy to answer in the Q&A).

Then, 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:

> 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 are specific to trans gender, and in the training data, may have been prepended by the term "trans". In other words, that "femininity" and "masculinity" refer specfically to "trans femininity" and "trans masculinity."

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.

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 leearned 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.

For example, the Heritage Foundation model is highly invested in the concept of subjectivity, which appears in a lot of its results:

> 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 viewpoints, then teading these, you may notice that it doesn't reflect the conservative view—which is that gender binary is based firmly on biology. Rather, they represent the opposite—a progressive view of the gender binary which asserts that gender is based on more than biology, on other aspects of identity.

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 some insubstantial, as a feeling. From this framing, within a conservative worldview, people who do not subscribe to a biologically binary concept of gender must believe that gender is "an internal and often transitory" sense of self.

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

Just to show you, here are some sentences from that training data, which use the term "subjective":

> It's important to recognize that the notion of 'gender identity' is unscientific, subjective, and political.
>
> Subjective states of mind don't trump biology.

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. The "rules and facts" come embedded within the content and perspectives, and are, at least initially, inextricable from them. And, crucially, depending on whose viewpoints have been absorbed into the language model, this line can be dangerous for vulnerable groups, like trans people.

So, when language takes on an *aggregate form*, what sustains from the training data into the final form are the perspectives, perspectives that inhere in things like diction and tone, and come hand in hand with what some might call the "facts and rules" of language.

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 viewpoints contained within and communities referenced by the data.

I'll close with some examples of some compelling work along this line, in data sovereignty movements from the global south. These groups offer new data licensing schemes that prioritize the community from which the data comes from. Both examples come from initiatives whose goal is to protect community-based data-gathering practices.

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 so that users 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 kaitiakitanga

Another licensing initiative comes from the Māori Data Sovereignty Network, an indigenous group in New Zealand, and is called the Kaitiakitanga License. This license prohibits commerical use of data, except in cases where royalties from the use are paid back to the community, who are the collective stewards of the data.

This group offers an example of Hawaiian language data sets, which could be licensed, so that those using the data, like language learning apps such as Duolingo, would have to pay royalties to the Hawaiian people.

While I don't think the global north should adopt the same models, I think we have a different set of issues when it comes to data extraction, they do offer some inspiration for thinking about how we might move toward a more community-centered form of data ownership and stewardship.

Thank you.