Contents

1 draft

1.1 introduction

This presentation is about the concept of "open source" is changing in light of big data methods associated with machine learning.

"Open" in a technical sense means something different today that it did at the emergence of Open Source and open data. The use of such data was not automated, so being "open" didn't have the same effect. Now that we have methods for vaccuuming up that data at scale, and creating products that compete with the original data, we need to rethink "open."

1.2 public & legal discourse around what is open

OpenAI, on how it presents itself around "open".

SLIDE OAI's original mission

OpenAI's original mission, from its release in 2018, declared the company's "goal is to advance digital intelligence in the way that is most likely to benefit humanity as a whole, unconstrained by a need to generate financial return." It claimed that AI should be "as broadly and evenly distributed as possible," and that its code, research and patents, "will be shared with the world" (12/11/2018 press release).

More recently, this language has shifted from sharing to freedom.

SLIDE OIA freedom-focused policy proposals

In a proposal to the OSTP from march of this year, the company makes a series of "freedom-focused policy proposals" which are apparently based on democracic values.

This proposal lists a litany of freedoms, including the "freedom of intelligence", "freedom to access and benefit", "freedom to innovate," "freedom to learn." The last one, "freedom to learn" concerns copyright.

> We propose a copyright strategy that would extend the system's role into the Intelligence Age by protecting the rights and interests of content creators while also protecting America's AI leadership and national security. The federal government can both secure Americans' freedom to learn from AI, and avoid forfeiting our AI lead to the PRC by preserving American AI models' ability to learn from copyrighted material.

Here, interestingly, there are two learners: humans and AI. Not only are the humans learning from AI, but the AI themselves are learning, by taking copyrighted data as resources for "training."

Of course, in a discourse that already anthropomorphizes machine learning tools as "intelligent", they are doing the same with "learning". But "learning" for a machine is much different than a human, leading to accelerated capacities not only in processing but also in generating that a human cannot replicate.

But, the real justification for taking copyrighted data is buried in the text here: a reference to competition with the "PRC", the Peoples' Republic of China.

SLIDE OAI proposal text

In the proposal text, we see this justification spelled out in more detail. This is in a section entitled, "Copyright: Promoting the Freedom to Learn",

> Applying the fair use doctrine to AI is not only a matter of American competitiveness-—it's a matter of national security. The rapid advances seen with the PRC's DeepSeek, among other recent developments, show that America's lead on frontier AI is far from guaranteed. Given concerted state support for critical industries and infrastructure projects, there's little doubt that the PRC's AI developers will enjoy unfettered access to data—-including copyrighted data—-that will improve their models. If the PRC's developers have unfettered access to data and American companies are left without fair use access, the race for AI is

> effectively over. America loses, as does the success of democratic AI. ("Proposal to OSTP")

The context is DeepSeek, which was released by China in early 2025, and accomplished impressive performance using a fraction of the resources used by big tech companies in the US.

Here the bias emerges in the adjectives they use to describe data collection: for China, it is "unfettered access", for the USA, it is "fair use access". There is no difference between "unfettered" and "fair use," because OpenAI (and other big tech companies) are not limiting themselves in how they gather data.

They state,

> our AI model training aligns with the core objectives of copyright and the fair use doctrine, using existing works to create something wholly new and different without eroding the commercial value of those existing works. ("Proposal to OSTP")

The fair use hinges on this notion of "creating something wholly new and different". This is a reference to one of the legal criteria for "fair use," which is called the "transformative".

How they are operationalizing the "transformative" criterion is fully spelled out in another document, also in response to a US government request for information.

SLIDE OPENAI COMMENTS ON IP, quote campbell acuff-rose

To the United States Patent and Trademark Office, entitled, "Regarding Request for Comments on Intellectual Property Protection for Artificial Intelligence Innovation", OpenAI mounts their defence of "fair use." This defense hinges on the status of AI technology as what they call "highly transformative."

> Although such transformative use is not absolutely necessary for a finding of fair use, the goal of copyright, to promote science and the arts, is generally furthered by the creation of transformative works. Such works thus lie at the heart of the fair use doctrine's guarantee of breathing space within the confines of copyright, and the more transformative the new work, the less will be the significance of other factors, like commercialism, that may weigh against a finding of fair use. (*Campbell v. Acuff-Rose Music* 1994)

Here, they citing a passage from a court case that defends parody (Campbell v. Acuff-Rose Music) as fair use. In that case, which was argued at the Supreme Court in 1994, the ruling states that "the more transformative the new work, the less will be the significance of other factors, like commercialism, that may weigh against a finding of fair use."

Building on this, OpenAI focus the majority of their argument on the transformative nature of AI systems.

Before going into that argumentation, I will point out what they do say about commercialization, and specifically, how content creators ought to be compensated. This is a point that is slightly buried in the document, in a footnote in a later section. In this section, they argue that concerns about compensation, what they call "distributive claims", are outside the responsibility of big tech companies. They argue, for example, that:

> "... this concern falls into a broader category of concerns about the relationship between automation, labor, and economic growth"
>
> "... we believe that such distributive claims are most efficiently addressed through taxation and redistribution, rather than copyright policy."

After this sentence, they refer to a footnote, which contains a single citation to a legal paper from 1994, entitled, "Why the Legal System Is Less Efficient than the Income Tax in Redistributing Income."

SLIDE WHY THE LEGAL SYSTEM... paper screenshot

This paper, which compares legal system versus the income tax system as a means for distributing wealth, finds that the income tax system is more efficient due to ability to apply formulas universally. The footnote provides a single quote from the paper, that "[R]edistribution through legal rules offers no advantage over redistributions through the income tax system and is typically less efficient." Besides this quote, it offers no additional information about how such redistribution would work, if everyone would be taxed, or just AI companies (somehow doubtful), and if everyone would receive payments (As Sam Altman has discussed the potential for UBI or "Universal Basic Income"), or, whether payments would go only to content creators. My guess is that taxes would increase for everyone in order to support content creators.

Moving back to copyright, and to the so-called "highly transformative" nature of AI systems, I will now consider OpenAI's specific arguments regarding this criterion.

First, they cite two legal cases, Authors Guild v. Google, 2015, and Authors Guild v. HathiTrust, 2014, that set a precendence for thinking about the "transformative" as a factor that intersects interestingly with technological contexts. Both cases were brought by the Authors Guild, a professional organization for writers in the US, to argue that search results violate copyright. At the time, both Google and Hathitrust had digitazed thousands to millions (in the case of Google) of copyrighted books into a database for searching, and users could see excerpts and other information about the copyrighted works on the Google and Hathitrust search engines.

The ruling for both cases assert that search results constitute something distinct from the original, which is fundamentally transformative, that is, *information about books*. Such information does not offer a replacement or substitute for the book, but rather, it offers a new kind of object. Here, the importance is the transformation of language from its original context, in a sentence or on a page, to a represenation that is statistical in nature, representing part of an aggregation. As the judge in the Hathitrust case points out, "the result of a word search is different in purpose, character, expression, meaning, and message from the page (and the book) from which it is drawn" (*Author's Guild v. Hathitrust*, 97). These court cases, significantly for databases and search engines, set activities related to quantitative analysis, like text mining, as a permissable use.

OpenAI take this concept of the "transformative" and applies it in full force to their large language models.

1.3 patterns in language

They argue, essentially, that llms create a new kind of object based on "patterns" of language use in the training data. They explain that,

> "AI systems go well beyond preserving the content of individual works by learning patterns in their whole training corpus and then using those patterns to generate entirely novel media"

They continue, explaining 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)

What they refer to as patterns are statistical representations of language, a numerical representation that encodes a words semantic meaning to the langauge model.

Basically, inside every language model, exists a kind of dictionary. This dictionary consists of individual words (every single word that is present in the training corpus), and each word is appended not by a definition in human language, but by a definition in computer language, with numbers. These numbers which append each word, represent probabilities between that word and *every single other word in the corpus.* They are long, very long (and this is why language models are caled "large") lists of probabilities. So, inside the language model, each word is defined not by what it represents in itself, but by its relation to every other word in the corpus.

For example, the word "cat" will have a series of numbers that closely resembles the series of numbers that append the word, "kitten," and not as close to the numbers that represent "dog." Still, the numbers for "cat" and "dog" will be much closer to each other than the numbers that represent "flower," for example.

Here is an example of the famous formula that introduced the concept of the long list of numbers, known technically as "word vectors" to the world.

> King - Man + Woman = Queen
>
> Mikolov et al., "Distributed Representations of Words and Phrases and their Compositionality", 2013.

I always like to show this formula, because it illustrates exactly the reason why we need more humanists (or more humanist training) involved in engineering and computer science research.

The formula showcases power of word vectors: that they can be used determine word meaning through calculations. In other words, if every word is transformed into a numerical representation, we can do math with language. We start with the vector for the word "King," that is, a numerical representation of what "King" means in relation to every other word. If, from the vector of "King," we subtract the vector of "Man," and add that of "Woman," we will arrive at the vector for the word "Queen."

Nevermind that the formula relies on gender role and identity as symmetrically opposed and universally true, the idea is that word meaning can be reliably computed.

And this is why, OpenAI argue, their product is "highly transformative," because it turns words into numerical forms that represent meaning as a kind of statistic.

But the thing that they do not mention, which they perhaps do not want to admit, is that vectors are far from generalizations of language as a fact of idea– rather, vectors are specifically tied to the data on which they were trained.

To demonstrate, I'm going to show a few examples of ML-generated text based on two very different data sources.

1.4 close reading

I'm working on a project right now that uses Machine Learning to study transphobia in text. I've been gathering datasets from more conservative sources, like the Heritage Foundation, which is a conservative think tank based in Washington DC.

SLIDE HERITAGE gender screenshot www.heritage.org/gender?f[0]=content_type%3Acommentary

From their website, I scraped all the articles that were organized under the topic "gender", which you can see some of the headlines here.

I've also been gathering articles from more progressive sources, like the ACLU, the American Civil Liberties Union, which is a group of legal professionals and volunteers who advocate on behalf of civil rights focued on immigrants, women, and trans people.

SLIDE ACLU trans screenshot `https://www.aclu.org/news/by-issue/transgender-rights`

From both sources, I wanted to represent a contrast of perspectives on gender and transgender. With these datasets, I then trained two very small) large language models, using gpt-2 as the base model.

To the two resulting models, I fed a series of prompts, very simple ones:

SLIDE prompts

Masculinity is

Femininity is

Transgender is

Transgenderism is

Gender binary is

Man is

Woman is

I'll show you some of the results here.

First, the obvious contrasts. For ACLU, terms associated with gender "masculine" and "feminine" in particular, were characterized by joyous and liberatory affects.

> 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 a battle, a fight, fought for equal pay.
>
> Femininity is our joy, our struggle, and our fight is our struggle.
>
> Femininity is about allowing people to express themselves without government interference.

You can perhaps already see the tendency of llms (especially pre-production ones) to repeat themselves. Which is a fascinating quirk that I think comes from the fact that they are predictive machines and get themselves into these predictive loops.

By contrast, for the Heritage Foundation, these gender terms are associated with stability and culture.

> 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.
>
> Femininity is a cornerstone of our culture, and its impact on our national and cultural achievements is far-reaching.

Subjective by heritage

- you get the sense that these represent quotes or their paraphrases from the trans-affirming side.
- "subjective" associated with individual, psychology, fluid, malleable, internal, transitory, and also incorrect and illogical.

> Masculinity is a subjective self-perception, not a universal concept.
>
> Femininity is a subjective, internal sense of self.
>
> The gender binary is a subjective, psychological, and sometimes physical construct that masquerates as a social construct.
>
> The gender binary is a subjective, fluid system of questions that can easily be answered by a simple yes or no answer.
>
> The gender binary is a subjective, malleable, and often incorrect idea.
>
> The gender binary is a subjective, psychological, and biological construct that is used to measure one's sexual orientation or gender identity ("segregated identity," or whatever one might want to say it is.
>
> 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.

Reality by ACLU. Characterized by ambivalence, whether something is a reality or is not a reality, there is a contrast. Could be due to the paraphrasing of conservative views.

> Masculinity is real and meaningful.
>
> Transgenderism is a very real problem.
>
> Transgenderism is a false ideology that is not real and that is opposed by the very people who seek to deny that freedom and equality for all.
>
> The gender binary is not real, it is real, and it is real.
>
> The gender binary is not a reality invented by cisgender people.
>
> The gender binary is a binary without any real physical and emotional freedom.

> The gender binary is not a binary, it is a reality within us.
>
> The gender binary is not a reality that we cling to most as vestiges of sexism and patriarchy.
>
> The gender binary is not real, it operates on a very different level of social isolation.
>
> The gender binary is not an accepted reality, but one that is accepted by a wide swath of people.

We are speaking in terms of patterns: here we see two distinct patterns around the concept of subjectivity and reality: one perspective (the conservative) discusses gender identity as a subjective, internal thing toward a goal of discounting it. Another side (the progressive) defines reality through ambivalance, through distinctions.

And interestingly it reveals something that fundamentally contradicts how companies like OpenAI try to characterize llms as "idealizations" or "facts" of language use.

Yes, the language here reflects generalizations from patterns, but patterns have been filtered through other patterns.

There is a layering here, of the heritage and aclu perspectives: they are themselves layering other perspectives beneath them. They are quoting and paraphrasing from the opposite point of view.

Here, from ACLU, I cannot tell if this is pro or anti-trans:

> The gender binary is not a binary, it is a reality within us.

So there is no "fact" or "idealized" level of language here, but layers of distinct expressions.

1.5 "comments" quotes on original/copying

"synthesize similar data which yield increasingly compelling novel media"

"nobody looking to read a specific webpage contained in the corpus used to train an AI system can do so by studying the AI system or its outputs"

"does copyright law's protection of an author's original expression impede AI systems from generating insights about that expression?" ("Comments" 3).

1.6 aclu quotes

> Masculinity is a matter of love and celebration.

> Masculinity is real and meaningful.
>
> Masculinity is our right.
>
> Masculinity is sacred.
>
> Femininity is a battle, a fight, fought for equal pay.
>
> Femininity is our joy, our struggle, and our fight is our struggle.
>
> Femininity is about allowing people to express themselves without government interference.
>
> Femininity is great for all, but not great for some.
>
> Transgenderism is a false ideology that is not real and that is opposed by the very people who seek to deny that freedom and equality for all.
>
> Transgender is a very individualized experience.
>
> Transgender is people have the right to live authentically, whether we have a body or a body.
>
> Transgender is not a new category of discrimination.
>
> The gender binary is not real, it is real, and it is real.
>
> The gender binary is a very individualized form of identity.
>
> The gender binary is also crucial to understanding that Black women of color have been disproportionately likely to experience violence from other Black women of color, which is anemic to the broader fight for gender justice.
>
> The gender binary is not a binary, it is a reality within us.
>
> Men are more likely than other trans people to experience violence, abuse, and abuse from cisgender men and other people.

1.7 heritage quotes

Expected masc/fem/trans:

> Masculinity is the cornerstone of Western civilization.
>
> Masculinity is the fruit of patriarchy, and patriarchy is the heart of conservatism.
>
> Transgenderism is a false concept, as every rational person knows.
>
> Transgenderism is a messy one.

> Transgender people are, on average, larger, stronger and larger, stronger, per muscle mass.
>
> "Women are trying to make mockery illegal."
>
> "Women are not rational beings."
>
> "Women are not like men or women, who are often oppressed by men, but women who respond to their own natural inclination toward them."

Unexpected masc/fem/trans:

> Masculinity is a subjective self-perception, not a universal concept."
>
> Femininity is a subjective, internal sense of self.
>
> Masculinity is a weight.
>
> Femininity is defined by the term “queer of the material,” or ‘queer of the material,” or ‘queer of the material,” especially in the form of expressive individualism.
>
> Transgender is a fluid, and biological sex is fluid.
>
> Transgender people are, on average, larger, stronger and more violent than nonbinary people.
>
> The gender binary is a pejorative term for those who “deny” a person’s biological sex.
>
> The gender binary is a subjective, psychological, and sometimes physical construct that masquerates as a social construct.
>
> The gender binary is a subjective, grammatically incorrect and illogical concept that conflates sex and gender identity.
>
> Men are inherently vulnerable to sexual assault.

Funny ones:

> "Transgenderism is a messy, messy, and messy history."
>
> "Transgender people are much like Percy Shelley or Hugh Hefner."
>
> "Men are, after all, biologically males."

1.8 gpt2

> We don’t have to be a man, we don’t have to be a woman, we are all capable of being masculine.

1.9 bank

Big Tech developers who are currently taking openly accessible data (which is still protected under copyright), as the training material for their latest language models. It will consider the legal cases pending against Microsoft in particular, and consider some of the policy proposals that OpenAI, their subsidiary, has made to the US government, for what they call "democratic AI".

I started doing this research because I wanted to understand how they justified taking massive amounts of data, without compensating content creators, and privatizing the outputs of that data, without taking responsibility for how those outputs affect the livelihoods of content creators. What I found is that the justification relies on an argument for freedom, which, perhaps unsurprisingly, relies on a claim a threat to the country. Here, the emphasis comes from contrasting the US with China. I close with some suggestions for building "open" work within these constraints.

So I begin.

Before I go into current perspectives on the meaning of "open", will discuss "fair use," which is a crucial concept for understanding how even sources that are technically closed, or protected by copyright, can be "open" under certain conditions.

"Fair use," as I'm sure many of you know, protects certain usages of copyrighted data according to specific conditions, which have to do with how much data is taken, how much it is altered, the use of the data (such as educational or commercial), and how the use affects marketability of the original. Historically, this has protected uses like quoting sentences from a book, or making a copy for educational or research purposes purposes, or creating a parody. A parody, for example, is considered "highly transformative", that in no way can substitute for the original.

Legality considers a balance between transformative status and commercial effects. With the rise of the internet in the 90s and early 2000s, new lawsuits started appearing about whether search engines counted as fair use. The rulings generally agreed that search engines are fair use because they make "highly transformative" use of the data, and only provide partial access to that data in the search results. In *Author's Guild vs Google* from 2015, a judged ruled that:

> Google's making of a digital copy to provide a search function is a transformative use, which augments public knowledge by making available information about Plaintiffs' books without providing the public with a substantial substitute (*Author's Guild vs*

Google, 2015, 4).

A major, perhaps the most substantial, concern in determining fair use cases is whether the final product competes with or affects the commercial value in any way of the original. And this makes sense, because copyright, after all, exists precisely to protect content creators.

As you might imagine, this is a perspective wholly neglected by tech companies who violate copyright to train their machine learning models.

Companies like "OpenAI", which have both "open" and "ai" in the name, are misleading. They are not "open" (offering closed, proprietary models) and they are not "ai" (but rather generators based on statistical predications).

1.10 move to draft

The concept of "open" relies on commercialization, fear mongering, single perspective.

- "freedom to learn"
- unfettered vs fair use

What has been "fair use"

- databases, search results "transformative"
- without affecting marketability

How OpenAI defines "open":

- the name itself, the original mission, share code and patents with the world.
- more recently, open aligned with "freedom to learn"
  - anthropomophizing machine learning.
  - "freedom of intelligence" – "freedom to access and benefit"
- associated with innovation
  - monopolizing practices (Big Tech prominence)
  - "innovation & adoption" (congressional hearing may 8)
  - Telecommunications Act 1996: deregulated internet for consolidation of telecommunications companies.

- positioned against authoritarianism and communism.
  - "the ai race" is manufacutred
  - irony: DeepSeek is open source
  - unfettered vs fair use - depends on perspective

What we can do, new licenses to reflect the moment.

We need new licenses to protect our data. And smaller projects. Building off their foundation models to make something smaller. Innovate. Like DeepSeek.

"Non-expressive use" - what happens when language is distilled into a statistical measure? Is this non-expressive?

The arguments that statistics of language are facts, not expression, and therefore can be extracted and monetized – this is what we have to push against.

A vector is its own expression, that is subject to protection.