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.

1.2 writing notes

1.2.1 ACLU close reading

From the other side, the ACLU-generated text, we see the exact same phenomenon. Instead of "subjective", however, the investment is in the term "reality."

> Masculinity is real and meaningful.
>
> 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 binary, it is a reality within us.
>
> The gender binary is not an accepted reality, but one that is accepted by a wide swath of people.

Here there is more ambivalence around the term "real", which depends on whether it is being asserted in a positive way, such as "Masculinity is real," or within a negative construction, such as "Transgenderism is a false ideology that is not real."

In a couple of the examples, this ambivalence is directly at odds. For example, in this amusing construction, "The gender binary is not real, it is real, and it is real."

1.2.2 "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).

1.2.3 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.

1.2.4 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."

1.2.5 gpt2

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

1.2.6 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. 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. commericalization 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.

2. fair use, campbell case

> 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.

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.

3. word vectors 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.

1.2.7 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.

1.3 reading notes

1.3.1 Chandrasekhar 2025

- how do copyleft licenses transfer to datasets, models, tokens?
- EleutherAI developing the Pile V2
- Problem isn't that data is used without compensation, but that products/outcomes are not contributed back to the commons (19).
- are parameter's "transformative"?
- The issue becomes: who has the ability to create? To use the GPUs.

- alternative licenses:
  - Nwulite Odobo "dual regime" - free for users in developing countries, multiple licensors for a dataset
  - Kaitiakitanga - royalties go to community, community ownership
- language is extractive, indigenous communities know this.

1.3.2 The Author's Guild v. Hathitrust, 2014

"A district court ruled that libraries that provided a search engine company (Google) with books to scan were protected by fair use when the libraries later used the resulting digital scans for three purposes: preservation, a full-text search engine, and electronic access for disabled patrons who could not read the print versions. On appeal, the Second Circuit affirmed fair use as to the full-text database ("a quintessentially transformative use") and as to use of text in formats accessible to print-disabled people (although not a transformative use, it is still considered a fair use based on the Betamax decision), but remanded the issue of fair use for long-term preservation of books." ("Summaries of Fair Use Cases", Standford Libraries)

1.3.3 Authors Guild v. Google, Inc., No. 13-4829 (2d Cir. 2015)

"Google made digital copies of millions of books submitted to it by libraries, scanned them and made them available to search through its Google Books service, so that users could—for free—identify relevant words, terms, or snippets from the scanned text. Google also allowed participating libraries to retain the copies they submitted. Important factors: Google's digitization was deemed a transformative use because it provided limited information about the books without allowing users more complete access to the works." ("Summaries of Fair Use Cases", Standford Libraries)

1.3.4 "Winning the AI Race: Strengthening US Capabilities in Computing and Innovation.Sam Altman, Testimony, May 8:

- May 8 congressional hearing titled "Winning the AI Race: Strengthening US Capabilities in Computing and Innovation."
- OpenAI CEO Sam Altman, Microsoft President Brad Smith, AMD CEO Dr. Lisa Su, and CoreWeave CEO Michael Intrator speaking to the Senate Commerce Committee.

- Argument: that the US requires free rein (low regulation) to defeat China in the "AI Race", we will know we win the race if we can innovate and export"
- Cruz:
  - position: regulation is "needless" and "orwellian", "paternalistic".
  - Cruz's contradictory language frames US as free, Europe and China as authoritarian.
    * Cruz's language contrasts "entrepreneurial freedom and technological innovation" against "command-and-control policies of Europe".
  - Drawing from history of the internet, which was developed with relatively low regulation in the USA.
    * Telecommunications Act of 1996 that promoted competition via deregulation, (but in reality, smoothed the road for consolidation, "going against its very stated intention by indirectly restricting newcomer access to broadcasting" (wikipedia, "Telecommunications Act of 1996")
  - Referring to Biden and some state legislatures: "They want a testing regime... seemingly something out of Orwell ... as if AI engineers lack the intelligence to responsibly build AI without the bureaucrats"
  - "U.S. dominance in AI depends on two factors: innovation and adoption."
- Altman:
  - vetting systems would be "disastrous" for industry, "sensible regulation that does not slow us down"
- Smith, microsoft president:
  - the way to know we've won the "race" is if our tech is broadly adopted.

1.3.5 NYTimes complaint

- NYT complaint argues that OpenAI "stea[s] audiences away from it", that outputs "compete", "closely mimic" NYT articles, and that the work is not "transformative". (page 4).

–> argumment seems to be about outputs being copies, when should be about inputs?

1.3.6 2018 OpenAI press release, december 12 2018, "Introducing OpenAI"

- OpenAI started as a nonprofit, and raised money with promises to share their products freely:
  - "Researchers will be strongly encouraged to publish their work, whether as papers, blog posts, or code, and our patents (if any) will be shared with the world" (OpenAI 12/11/2018 press release).

1.3.7 2025 "OpenAI's proposals for the U.S. AI Action Plan" march 13, 2025

- "we must ensure that people have freedom of intelligence, by which we mean the freedom to access and benefit from AI as it advances"
- "freedom-focused policy proposals"
- "neutralizes potential PRC benefit from American AI companies having to comply with overly burdensome state laws."
  - "freedom to innovate" regulations
  - "copyright strategy that promotes the freedom to learn"
    - "secure Americans' freedom to learn from AI"
    - "avoid forfeiting our AI lead to the PRC"
  - "export strategy"
  - develop infrastructure
  - adoption by government

1.3.8 2025 OSTP OSTP proposal, march 13, 2025

- Office of Science and Technology Policy proposal
- proposals to help OSTP develop "AI Action Plan ... that ensure[s] that American-led AI built on democratic principles continues to prevail over CCP-build autocratic, authoritarian AI".
- "democratic AI"

  - "a free market promoting free and fair competition.
  - "freedom for developers and users to work with our tools"
  - "preventing government use... to amass power and control their citizens"

- Deepseek is a threat because "simultaneously state-subsidized, state-controlled, and fully available... cost[ing] users privacy and security."

- point #3: "Copyright: Promoting the Freedom to Learn"
  - need to use copyrighted material to compete with China, a "matter of national security."
  - contradiction between China's "unfettered access" vs OpenAI's "fair use":
    * "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. Ultimately, access to more data from the widest possible range of sources will ensure more access to more powerful innovations that deliver even more knowledge" (10-11).

1.3.9 2023(?) OpenAI Comments on Intellectual Property Protection for Artificial Intelligence Innovation

- argue that, "Under current law, training AI systems constitutes fair use"
  - argument for fair use hinges on "transformative" use of copyrighted work
    * citing a passage from a court case that defends parody (Campbell v. Acuff-Rose Music) as fair use to argue that AI outputs are "highly transformative"

        * input data: copyrighted works become statistical patterns, "non-expressive""
        * output data: nobody can use AI to read the specific webpages they are trained on: they will still go to NYTimes to read the news. (debatable).

- "mission is to ensure that artificial general intelligence ("AGI") benefits all of humanity""
- anthropomorphize AI training into human learning:
    - "does copyright law's protection of an author's original expression impede AI systems from generating insights about that expression?"
    - ""training" refers to the process by which an AI model learns patterns"
- "Authors may object that the outputs of generative AI systems will harm the value of their works. We address this objection in Section II."
    - "Distributive Issues from AI-Generated Non-Infringing Works Should Be Addressed by Other Policies"
    - "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."
        * "Louis Kaplow & Steven Shavell, Why the Legal System Is Less Efficient than the Income Tax in Redistributing Income, 23 J. Legal Stud. 667 (1994) ("[R]edistribution through legal rules offers no advantage over redistributions through the income tax system and is typically less efficient.")."