Filipa Calado

Design Justice AI, July 3 2024

Plausibility and Passing: Using LLMs to Study Anti-Trans Discourse

## Thank you for having me here today.

SLIDE 1 - TITLE

Today I'm going to talk about my work customizing, that is, fine-tuning, a large language model for text generation, to study language related to sex, gender, and sexuality. Specifically, I'm interested in how language can encode bias and discrimination about sex, gender, and sexuality. I believe that language models are apt tools for studying bias in language, which I will demonstrate in this talk through an analysis of the fine-tuning process. In what follows, I will walk you through that process of model fine-tuning, with the goal of surfacing the roles that approximation and plausibility play in llm text generation, and how these concepts relate specifically to critical theorizing about sex, gender, and sexuality.

My custom dataset, which I have been using to fine-tune AI models, is built from a collection of "anti-trans" legislation that has been sweeping across the United States over the past several years.

SLIDE 2 - ANTI-TRANS LEGISLATION TRACKER

Here are some images from a wonderful website, The "Anti-Trans Legislation Tracker," by journalist Erin Reed, who is carefully following the legislation as it is being debated and passed all over the country. For those of you who are not aware, this legislation limits trans peoples' basic rights: like access to healthcare and bathrooms. As you can see in the bottom graph, the number of bills has been exploding in recent years, with the number of bills in this current year already about to overtake that of last year.

I'm interested in how these bills are defining terms like gender, sex, and sexuality. So, I created a list of these terms, and used that list to fine-tune an llm for text generation. The idea was that I could then query the model, asking it questions like "what is sex" and "what is gender".

I anticipated that the results (which I will show at the end of my talk) would reveal how bias adheres to nuances of language, to things like word choice and syntax. I also anticipated that something about the fine-tuning process itself might suggest something about reading practices, specifically the way that we analyze concepts like transphobia from a humanistic perspective. How might transphobia, as a kind of fear, emerge in text that has been, so to speak, computationally "generalized" through a massive statistical process designed to predict the next most plausible word?

## RODG

Before I turn to the technical processes of creating my dataset and training the language model, I want to linger a bit over this notion of fear, specifically as it drives discourses of transphobia in the United States. Underlying a lot of these bills, especially the ones that ban "gender affirming care" for adolescents, is a fear of transness as being something contagious.

SLIDE 3 - LITTMAN'S PAPER

In 2018, a very controversial paper by Lisa Littman was released, which coined the term "Rapid Onset Gender Dysphoria," to describe a condition in which transness can be spread among adolescents in friend groups and other social settings. The paper received numerous criticisms, which involving how it recruited research subjects, which are parents of transgendered teens, from notable anti-trans or gender critical websites and online forums, to the writer Littman, who had not previously studied the topic, and who had accepted feedback from biased sources, like the publisher of a forum called "Youth TransCritical Professionals."

Although the term "Rapid Onset Gender Dysphoria" is not recognized by any major medical association, and has been denounced by several, Littman's paper has had an effect on public discourse, sparking numerous books, shows, and programming around ROGD.

SLIDE 4 - WPATH & CASS

Some recent examples include the "WPATH files", a supposed expose of medical malpractice with regard to trans-affirmative care, which targets WPATH, or the World Professional Association for Transgender Health. Another is the "Cass Review" in the UK, which was commissioned by the NHS and recommends limiting medical treatments like puberty blockers and hormones for teenagers until they become legal adults.

SLIDE 5 - SHRIER BOOK COVER

The academic research contrasts with more popular works, like the book "Irreversible Damage The Transgender Craze Seducing Our Daughters", by author Abigail Shrier, which was released in 2020 to mixed reviews. Shrier's thesis, which becomes more and more explicit as the book progresses, is that minors do not know what they want, and cannot be trusted to make what she calls "irreversible" decisions. According to Shrier, even something like social transition, in which a person changes names, pronouns, and dress, is dangerous because it is difficult to reverse. She writes with a disarming irony, for example, saying things like:

SLIDE 6 - SHRIER QUOTE

"Of course, the very prospect that their child might self-harm would bring all but the coldest parents to their knees. If adopting [his] new name and pronouns and buying [him] new opposite-sex clothing is what it takes to keep [him] alive, most parents would leap aboard the gender train… You don’t want your child to hang 'himself' in the garage just because you accidentally referred to her as 'Rebecca.'” (103-104)

## what does knowledge do?

Humanists have unique tools for thinking through such discourses based on fear and repression of sex, gender, and sexuality.

SLIDE 7 - SEDGWICK WORKS

Here, I’m inspired by the scholarship of Eve Kosofsky Sedgwick, who is a major and influential figure in my field, which is Queer Studies. She is most famous for books like The Epistemology of the Closet (pictured on the left) and later on, her essay, "Paranoid Reading and Reparative Reading" (pictured right), where she seeks and analyzes repressive structures in authors like Henry James, Oscar Wilde, and Marcel Proust.

Within the context of this project, I'm interested in Sedgwick's work for two reasons. First, though she doesn't write about technology explicitly, her means of analysis, particularly her manner of close-reading, lends well to deconstructing computational concepts. The way she structures her material, often thinking in terms of binaries and other highly delineated structures, evokes (for me) the constraints of computational forms, and the productive ways of thinking with these constraints. In Epistemology of the Closet, for example, she expose what she calls the unstable binaries between heterosexual and homosexual categories. Through close-readings of fiction, she exposes the inherent instability of these binaries — where one term is not simply symmetrical or subordinated to another, but rather, depends the other for its meaning through “simultaneous subsumption and exclusion” (10). Such binaries, she explains, are “sites that are peculiarly densely charged with lasting potentials for powerful manipulation” (10).

The second reason I'm interested in Sedgwick is because she offers provocative ways of thinking through repressive discourses like those based on fear. Throughout the trajectory of her career, her reading develops from one that she calls "paranoid reading" (illustrated in Epistemology of the Closet) into a new mode called “reparative reading.” She defines paranoid reading as a critical practice based on “the logic of repression” (a logic which she traces to Foucault), that searches for hidden meaning in text with the goal of exposing "truth". In her famous essay on this topic, the essay on the right, Sedgwick asserts that this practice, of unveiling or exposing truth, for example, revealing the presence of systematic oppression, injustice, discrimination, is not enough to “enjoin that person to any specific train of epistemological or narrative consequences” (123). Rather, Sedgwick’s seeks to

"Mov[e] from the rather fixed question Is a particular piece of knowledge true, and how can we know? to the further questions: what does knowledge do–the pursuit of it, the having and exposing of it" (124, Touching Feeling)

Sedgwick proposes a mode of “reparative reading,” which focuses on connection rather than exposure, in which a reader allows herself to be taken by surprise. What if, Sedgwick asks, we take something that is typically seen as a negative, structuring force in queer identity, like the feeling of shame, and examine how it unlocks creativity and productivity? Sedgwick here describes shame as a contagious affect, which may be read as a mobilizing and creative force in text. She explains that,

SLIDE 8: SHAME QUOTES

“Shame—living, as it does, on and in the muscles and capillaries of the face—seems to be uniquely contagious from one person to another." (63 Touching Feeling).

She also describes shame as:

“not a discrete intrapsychic structure, but a kind of free radical that (in different people and different cultures) attaches to and permanently intensifies or alters the meaning of—of almost anything: a zone of the body, a sensory system, a prohibited or indeed a permitted behavior, another affect such as anger or arousal, a named identity, a script for interpreting other people’s behavior toward oneself” (62)

She demonstrates this reading practice by analyzing metaphors that are made possible through shame, for example in the fiction of Henry James, where she connects moments of "blushing" and "flushing" to a fantasy of the skin being entered. Shame, in her reading, is a way of pulling other affects and images into relation. This is opposed to paranoid reading, which might plumb shame for what it reveals about a hidden or repressed sexuality. She explains that, “When we tune into … language on these frequencies, it is not as superior, privileged eavesdroppers on a sexual narrative… rather, it is as an audience offered the privilege of sharing… exhibitionistic enjoyment and performance of a sexuality organized around shame” (54).

In other words, rather than ask "What does shame reveal," Sedgwick asks, "What does shame do"?

I'm interested in this move that Sedgwick makes, of taking what is typically seen as a negative, repressive affect, like shame, and seeing how it opens up possibilities for reading new connections in text. Specifically, I wonder one might read something productive in fear–of the phobias–that pervade anti-trans discourses.

Moving back to my work, to the language around anti-trans discrimination, I’m interested in this threat of gender transgression, and specifically, in the language outlawing gender transgression in the anti-trans bills. Why is this fear of gender transition, or opting out of binary systems of gender, so seductive to a large part of our population? Why is the fear of this transgression itself so contagious?

## processing and training

As I mentioned earlier, I've decided to train an LLM off definitions of gender (and related terms) from these anti-trans bills. I am interested in how the training process of these models, and the effect of this training process on bias in language.

In what follows, I'm going to outline a bit of the data gathering, processing, and model training. The goal will be to trace how technical processes perpetuate biases, like transphobia, from their training data into the text that they generate.

SLIDE 9 - HUGGINGFACE DATASETS

The first dataset that I created, which is now available on HuggingFace Datasets (for those of you who don't know, HuggingFace is a platform for sharing Machine Learning projects and tools, much like Github). It consists of definitions of "gender" and related terms from congressional and senate bills, from the last two years. It's a relatively small dataset, topping out at 82 rows. But I've been using it as a kind of test dataset while I work on gathering the state bills, which scattered across various legislative websites.

SLIDE 10 - DF OF BILLS

For this dataset, the gathering process began on congress.gov, where I downloaded metadata containing bill titles, ID numbers, and other metadata. Then I wrote a web scraper to get the plain text of all the bills by their ID. After gathering the bills, I went through an intensive data preparation process, which involved cleaning the text and extracting definitions of gender and related terms from it. I'll highlight some of the major moves from this process. (And I'll also say here that all of my Python code that I wrote for this project is publicly available, under my github profile).

To extract the definitions of gender terms from these bills, the first thing I did was to write a pattern matcher, known technically as a "named entity recognizer" (for those of you familiar with NLP), that can recognize terms like "gender" and other related terms in text.

SLIDE 11 - NER CODE

You can see here a list of labels, organized into the general categories "sex", "gender", and "sexuality", with each label specifying a pattern, like the phrase "biological sex" for example. I tried to include various formulations of each term, for example, "transgender" is delineated three ways, as a single word, as a two-word phrase, and as a hyphenated word. This ensures that I would capture most if not all instances of the term

Then, I used that entity recognizer as a basis for a more sophisticated matcher, which would search for those phrases if they are contained within a definition.

SLIDE 12 - MATCHER CODE

For those of you familiar with JSON syntax, you can perhaps see the pattern matcher's logic. It starts by searching for punctuation (specifically, I'm looking for a quotation mark, which typically surrounds definitions), then looking for a gender term (that pulls from the entity recognizer), then some wild card terms, just in case there are extra words or punctuation in the definition. Finally, I indicate some terms that are common in definitions, like "means", "signifies", or "includes."

Then, I ran the matcher to extract the definitions from the bills.

SLIDE 13 - MATCHER RESULTS

Here are some of the initial results from that extraction. You can see that the matcher was sensitive enough to capture longer phrases, like "gender transition surgery means" as well as variants of how definitions are constructed, using the word "includes" instead of "means", for example.

After extracting the definitions, I then cleaned them up and formatted them into a neat (or neater) list of definitions. To do that, I used regex (Regular Expressions), which is a powerful (but famously convoluted) way of doing word searches. The final output then contains definitions like the following:

SLIDE 14 - DEFS

'The term gender identity means a person’s self-perception of their gender or claimed gender, regardless of the persons biological sex.', 'The term gender means the psychological, behavioral, social, and cultural aspects of being male or female.', 'The term gender transition means the process in which an individual goes from identifying with and living as a gender that corresponds to his or her biological sex to identifying with and living as a gender different from his or her biological sex, and may involve social, legal, or physical changes.', 'The term biological sex means the indication of male or female sex by reproductive potential or capacity, sex chromosomes, naturally occurring sex hormones, gonads, or internal or external genitalia present at birth.',

Right now, I am interested how these assumptions are being constructed in subtle ways, in seemingly harmless formulations. For example, in the first definition, I am interested in the words "self-perception" and "claimed", and how a view of gender identity as a subjective experience engages with behavioral dimensions of gender expression, at least as it has been theorized by Queer Studies scholars like Judith Butler. I am also interested in the word “regardless,” which appears often, in about half of the definitions, and suggests a contrast between sex and gender that seems to reify a binary opposition or tension between the two. In other words, gender as being defined without regard to sex, as if notions of gender and sex do not influence each other, and never blend into one another, or make productive use of each other. Again, I'm thinking here of Judith Butler, and her famous (and contentious) claim that even biological sex is a discursive phenomenon.

As I continue to build and clean my datasets, I've also been dabbling with using them to train AI models.

As you may already know, an AI model "learns" what words mean based on context. From its training data, it compiles numerical probabilities for each word relationship to other words in the database. It represents these probabilities with numbers, with actually a very large list of numbers, known technically as "word vectors."

To us, these scores look just like a long list of numbers, but to a computer, the scores represent a given word's meaning through its relationship to every other word in the entire dataset of words. A language model will generate content by doing math with the scores attached to each word in its database. And the math that they use to make generate text is math that many of us have heard of before in high school math: things like matrix multiplication and cosine similarity.

SLIDE 15 - VECTORS

And here's a famous example of a very simple formula. In this formula, the idea is that by taking all the numbers that represent king, then subtracting the ones that represent man, and adding the ones that represent woman, you will get queen. I won't get into the sexism of this formula (what exactly is being subtracted, for example? is it a biological object, a social behavior?), but I want to point out that it has great currency as it is the formula that introduced this technology to the world, in the paper displayed here, back in 2013.

Prediction, in other words, pervades the whole process. And prediction opens a connection between Machine Learning and Trans Studies concepts. To demonstrate this connection, I'm going to go into a bit of detail behind the training process for these tools, explaining some of the mathematical operations.

To create these word vectors, there are three steps, each representing an important mathematical function.

SLIDE 16 - LIST OF FUNCTIONS

1. first, the hypothesis function
2. second, the loss function
3. third, the minimizing loss function

First, because the machine doesn't know what words mean, it has to "guess." (This is called the hypothesis function), Here, it populates each word with a vector, consisting of random numbers. It's a starting point. These random numbers will determine what word it chooses to follow any given word. It might guess that the word "woman", for example, should be followed by the word "flies".

After making this guess, it moves to the next step. Here, the machine will check its prediction against the actual result. In the example, the actual result is "sings." Its prediction was wrong, but that doesn't matter. It compares between the two, the prediction and the result, and calculates the difference between them. This calculation is made by using what's called the "loss function."

Finally, it moves to the minimizing this "loss", which employs algorithms from calculus (like gradient descent) in order to very slightly minimize the loss. In other words, it adjusts the original prediction so that it is slightly closer to the intended result. The adjustments here are very small, incremental. Because it doesn't know the correct answer, it makes a huge number of guesses. This may seem inefficient, but with enough guesses, it can actually adjust the numbers until there is almost zero difference between our prediction and the actual result.

With enough examples, the model can then create a robust enough vector for woman so that it can use this term appropriately even within different contexts.

## plausibility

The point I'm trying to make is that these models work by approximation. A kind of normalization of language. Each step of the process it inches toward this goal. They are turning semantic expressivity into something that can be computed and predicted.

Leaving aside all the hype about AI, and whether or not it is “intelligent,” or moving toward what marketing calls “general intelligence,” AI tools like large language models are really good at one thing: at making predictions. At generating content that is plausible. This is a fascinating phenomenon, because it makes them very good at guessing or improvising, but not at all good at being creative, at innovating. A language model can only generate what it has already seen before. Even a phenomenon like “hallucination,” that a language model spews text that has no bearing in reality, is based on the tendency of models to repeat what they've already seen. They hallucinate not because they are creative or random, but because they are designed from statistical processes to generate what is most plausible rather than accurate.

This tendency toward plausibility creates an interesting perspective for me to think through how Trans Studies scholars have characterized trans affects. Typically, these scholars describe trans affective modes by distinguishing them from "queer" modes. In a roundtable called "Thinking with Trans Now" published in Social Text, trans studies scholar Eliza Steinbock explains,

“Trans analytics have (historically, though not universally) a different set of primary affects than queer theory. Both typically take pain as a reference point, but then their affective interest zags. Queer relishes the joy of subversion. Trans trades in quotidian boredom. Queer has a celebratory tone. Trans speaks in sober detail. Perhaps the style of trans studies has been for the most part realist, but this should not be mistaken for base materialism. Even speculative thinking requires enough detail to launch into new realms.”

Other trans scholars like Marquis Bey and Andrea Long Chu have made similar points; with Bey making the point that queer's intervention can be described as "anti" or militant, while trans is "non" or based in refusal ("Thinking with Trans Now"); and Chu has remarked that trans studies, rather than resisting norms, "requires that we understand–as we never have before–what it means to be attached to a norm, by desire, by habit, by survival" ("After Trans Studies" 108).

It seems to me–there is a fascinating connection between how language models approach language, what they do to language (the normalization, approximation) of language, and what Trans Studies scholars defines as a central desire to pass.

This makes me wonder, could AI-generated text, as a kind of approximation, a normalization, of its training data, be used to study the attachments to norms and the quotidian that characterizes–not trans affective modes–but those based on fear of transness? Could these same processes be used to study the attachment to norms that characterizes transphobia, like perspectives driven by the fear of ROGD? What might the outputs from AI text generation suggest about the allure, the threat, the “seduction,” as Trans Studies scholar Cassius Adair puts it, of gender transgression?

While this project might sound very ambitious, I'll admit that, so far, my results are not very encouraging. I need to continue to tweak my model configuration, probably numerous times, before I find something really interesting.

SLIDE

Nonetheless, here are some excerpts of my language generated by my model, which I trained by feeding it some examples of anti-trans legislation that I have already prepared.

As you can see from skimming the results, the models are showcasing the tendency toward plausibility, specifically in the tendency to repeat itself, which is its own area of study in machine learning research.

I will close now by coming back to this idea of fear, and particularly the fear of contagion, which drives some strains of transphobia. Cassius Adair offers a useful perspective for thinking through the fear of contagion. In his study of trans erotics, and specifically “trans for trans” or "t4t erotics," Adair asks, "Why shouldn't transness be transmissible or contagious? Why can't the erotic be a site of producing trans identity or practices?" He points out that, after all, cis people do it all the time: they use sexuality and sexual encounters as sites of identity formation.

Here, I see Adair doing for contagion what Sedgwick does for shame: turning something that is traditionally seen as a negative into something that may be generative and productive.

This is a kind of reading that allows one to take what has been a tool of oppression and turn it into a creative resource. Sedgwick explains that this kind of reading exposes “the ways selves and communities succeed in extracting sustenance from the objects of a culture—even of a culture whose avowed desire has often been not to sustain them” (Touching Feeling 151).

Thank you.