1. Reading Fear: A meditation on methodology in machine learning experiments

1.1. thank you for having me

Thank you so much for having me here.

I'm going to talk about a project that I've been working on for a little over a year, which looks transphobia, and particularly anti-trans discourse, and the ways that it is currently proliferating across this country.

To study this topic, I'm using machine learning as my methodology.

I'm interested in how machine learning processes, in how predictive algorithms, which reconfigure and replicate the data used to train them, might be deployed to deliberately study bias and discrimination in language. For this project, I train a machine learning model off of data from anti-trans legislation, which is currently being proposed, debated, and passed all over the country, and which limits trans peoples' rights.

In this presentation, I'm going to engage my technical methodology with critical conversations in queer and trans studies. Afterwards, I am interested in hearing what you all think about this methodology from a humanist perspective. I'm wondering how my methodology conveys, and whether it is generative, for an audience that does not have a technical background.

1.2. legislation, transphobia, ROGD, littman

So to start. This project looks at transphobia within legal discourse, in the current anti-trans bills proliferating across the United States.

[SLIDE - FEDERAL BILL NAMES]

Here, you'll see a list of the most recent bill titles being proposed in Congress, like "Ensuring Military Readiness Act," and "Protect Children's Innocence Act." The majority of these bills target education, seeking to ban education on gender and sexuality, and students' chosen names and

pronouns, as well as healthcare, where many bills outlaw gender-affirming care for minors and adults.

I'll come back to these bill titles in a moment, but for now, it's enough to acknowledge that they are based on a certain logic that gender nonconformity is a national threat, from which the American people (in particular children) need protection.

This logic is driven by a particular fear of transness being contagious, that it can spread from person to person. This fear, which is popularized in the now debunked medical diagnosis called "Rapid-onset Gender Dysphoria," or ROGD, has had a significant influence on public perception around Trans issues, especially as they affect minors.

SLIDE - SHRIER BOOK COVER

One notable example of this is a book by Abigail Shrier, called "Irreversible Damage: The Transgender Craze Seducing Our Daughters." Shrier's thesis is that minors, who do not know what they want, cannot be trusted to make decisions about their gender.

Unlike the language of the legislative bills, Shrier's tone throughout the book is highly ironic, an irony that embeds some interesting subtexts.

SLIDE - SHRIER QUOTE

For example, there is a double meaning behind some of her pronouncements. She says things like: "if the government can’t force students to salute a flag, the government can’t force a healthcare worker to utter a particular pronoun. In America, the government can’t make people say things–not even for the sake of politeness. Not for any reason at all" (xx).

Here, Shrier makes a comparison to patriotism—in the flag salute—to make a point about governmental authority. The point is that, just as the government cannot compel people to express loyalty to the country's symbols, so it cannot compel respect (or here, she refers to it as "politeness") for a person's preferred gender designation.

Through this comparison to patriotism, and the framing of the issue around expression, the subtext here seems to be that forcing pronoun usage would be fascist. Which is a strange way to make a point to what (I imagine) is a largely conservative readership.

Unless, reading this statement through the frame of irony, the point is precisely that some kinds of expression should be free while others should not. That we *do* have the right to question each other's genders, but we shouldn't question the flag.

In plain terms, then, the sentence would be saying something like, "look, our government so loves your freedom, that it cannot even force you to show respect for it. If you are free to disrespect such a government, then you surely cannot be obligated to honor someone's pronouns".

In this reading, we can trace a direct line between an investment in patriotism to an investment to gender norms.

While I enjoy doing this kind of analysis of irony, it is exactly the kind of reading that I don't want to do for this project. Because the more that I engage with this material and the public discourse around it, the more I am convinced that this particular historical moment needs another mode of reading.

1.3. sedgwick, paranoia, shame, movement

Here I am inspired by a major figure in my own field, which is Queer Studies. As many of you know, one of the major figures of this field is Eve Kosofsky Sedgwick, who, throughout the trajectory of her career, explored various reading methods and modes of relation to text.

SLIDE - EPISTEMOLOGY

In her early work, Sedwick practices a mode of critical analysis that mines language forms for what is hidden, what is obscured in text, and what this reveals about binary structures and power relations. Then, in her later work, spurred by the horrors of the AIDS crisis and the lack of action by the government to address it, Sedgwick pursues ways of knowing not oriented around revelation

and exposure.

She asks, what if we "move 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).

To demonstrate, Sedgwick takes "shame," an affect that is traditionally seen as negative, and examines how it creates productive and generative effects in text. She describes shame as:

SLIDE - SHAME QUOTE

> “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)

I'm interested in this move that Sedgwick makes, of taking what is typically seen as a negative, repressive affect, and seeing how one might read something productive.

In my project, instead of focusing on what transphobia is afraid of, that is, the fear of gender nonconformity, what could I learn about its positive attachments? For example, what if we turned our attention to the desire for and attachment to normativity?

And this attachment to normativity, in fact, is one way that trans studies has distinguished itself with regard to queer studies, at least according to some scholars.

Trans studies scholar Eliza Steinbock explains that,

SLIDE 16 - TRANS AFFECTS

> “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.”

Similarly, Andrea Long 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).

You'll remember in the list of bill titles from before, the patriarchial undertones in words like "protect," "preserve" and "ensure." Within that language, the fear of change that they imply, there is also some kind of attachment to normativity, to maintaining tradition. It is that attachment that I'm interested in exploring.

Now, in the next section, I'm going to explain why I think that machine learning is a particularly good method for this task of studying normativity.

1.4. prediction

I'm going to go a bit into technical detail here, because the mechanism of the technology is important to my thinking through my method.

So, to put it most succinctly, the thing that interests me the most about machine learning is the way it works on prediction and plausibility. As many of you may know, all machine learning models (like the one that runs the ChatGPT, for example), make predictions, or guesses, as to what word should follow another word.

But how do they know what an individual word means? Here's the first complicated part: each word, in the model's "understanding," if we can call it that, is represented by a definition, a definition that consists of a long list of numbers. And these numbers, each of them, represent a very, very complex probability for that word's in relation to *every single other word*.

So, a single word is defined by, not what it means in itself, but how it relates to every single other word. (By the way, this is why the models are called "Large Language Models", they are large

because these lists of numbers are just massive).

Once a model has a list of numbers to represent each word, it can then use algorithms to calculate which words should be put together, side by side, in a sentence. In this way, text generation is really just turning language meaning, semantic expressivity, into something that can be computed with math, in numerical form. (And for those of you who want more information about this concept, which is known technically as “word vectors,” I am happy to explain in the Q&A).

And here's the second complicated part. To get these long lists of numbers, models must be trained. The training process can be reduced roughly into three steps.

SLIDE - LIST OF FUNCTIONS

1. hypothesis
2. loss
3. minimizing loss

The first step is the "hypothesis" step. Here, a model will take a sample sentence from the dataset, and it will block out the second half of that sentence. Then, it will try to guess which words should go in that second half. Because the model has no idea what the words mean, the guess will be wrong. But that's doesn't matter, because the purpose of the hypothesis is to make any guess, so that it has something from which it build on in the future steps.

Then, after making this guess, it moves to the next step, where the machine checks its prediction against the actual result—it will compare the predicted word against the actual word in the sentence. And it will calculate the mathematical difference between the prediction and the actual result, which is called the "loss".

Finally, in the third step, it moves to the minimizing this "loss" by *very slightly* adjusting the lists of numbers (attached to each word) so that they are closer to the intended result. The model will do this many times, making incremental changes each time, so progress is very slow, but also very

precise. (And this constant iteration of numbers, and the computer processing required to do it, is why language models take lots of time, energy, and computer hardware to train). At each round of training, the numbers attached to each of the words are slightly adjusted toward the most likely number, which is in effect, an average of that word's relationship to every other word in the database.

I read this iterative shifting of numbers (representing words) within the model as a kind of *approximation* or even *normalization* of language. The model generates language by approximating what is most likely based on its training data.

And this is exactly why, while models are good at guessing or predicting, they are not at all good at being creative, at innovating. A model can only generate what it has already seen before. Even a phenomenon like "hallucination," that a 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 most accurate.

1.5. plausibility

[SLIDE OF RESULTS]

Here are some of the results that I've gotten so far from my model training. As you can see, the results aren't so great right now. I'm still working on adjusting my model parameters to get more cohesive responses.

But so far, the preliminary results do suggest a certain repetition of language that bears out my point that plausibility that drives text generation. When the model doesn't know what to say, it just repeats what it already knows.

Here, I see a fascinating connection between how language models approach language, what they do to language (the normalization or approximation) of language, and what Trans Studies scholars

define as an attachment to normativity, that is, a desire to *pass*.

This makes me wonder, could generated text, as a kind of approximation, a normalization, of its training data, be used to study norms and attatchments to norms in the language that characterizes transphobia?

And if so, What might far-right investments in normativity illuminate about trans investments in normativity? What might we learn about the generative and creative effects of contagion? What might they suggest about the allure, the “seduction,” as trans studies scholar Cassius Adair puts it, of gender transgression?

Although this might be a controversial question, I think it's a necessary one in our current political moment. It is important to point out that Sedgwick's shift in reading methods was prompted by a sense of disillusion about the AIDS crisis, and the lack of action by the US government to protect the lives of those affected, which were predominantly black people and gay men. With the anti-trans moment, we are in a similar situation, I think, where the public discourse is so skewed, that we need new ways of thinking through fear and especially the fear of contagion.

I’ll finish with a final question by Cassius Adair. He 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.

SLIDE - THANKS AND CONTACT

Thank you.

And for those of you who want to look at the code and datasets I created for this project, you can find it on Github, under my username, *gofilipa*, and under the full link here at the bottom.