Filipa Calado, PhD

“Plausibility and Passing: Using Chatbots to Study Anti-Trans Discourse”

In the midst of the AIDS crisis in the US, Eve Kosofsky Sedgwick called for a critical reading method that aims beyond exposing the existence of oppression or discrimination. Even if critique could prove the government’s negligence (or even hatred) for those groups most affected by AIDS—Black people, gay men, and drug users, in particular—Sedgwick asks, “What would we know then that we don't already know?”[1]

Today, nearly 30 years after Sedgwick's landmark essay on “Paranoid Reading and Reparative Reading,” the brute fact of anti-LGBGTQIA+ discrimination surfaces again in wave after wave of legislative bills that limit trans peoples’ access to healthcare, public spaces, and more. This paper explores what new technologies might offer for the analysis of old biases. To study the discourse driving the contemporary anti-trans movement, I use a contemporary tool: a Large Language Model (LLM) chatbot. From hundreds of these anti-trans bills, I compiled a dataset to train and customize my LLM chatbot with the goal of studying transphobia within the language of the bills themselves. Specifically, I am interested in how terms like “sex,” “gender,” and “sexuality” are being defined in the bills, and how these definitions relate to the fear of transness as being something that spreads from person to person, like a contagion—popularized by the false phenomenon of “Rapid Onset Gender Dysphoria,” which has been denounced as false by major medical associations.[2]

Machine learning models, like the ones that power chatbots, operate based on prediction: they are “trained” to turn semantic expressivity into something that can be computed and guessed. At the start of the training process, the models guess what word should follow another word (using what is called the

[1] Sedgwick, Eve Kosofsky. “Paranoid Reading And Reparative Reading; Or, You’re So Paranoid, You Probably Think This Introduction Is About You.” *Novel Gazing: Queer Readings In Fiction* (1997): 4.

[2] This condition was coined by psychologist Lisa Littman in a highly controversial (and since retracted) 2018 paper. See Littman, Lisa. “Rapid-onset gender dysphoria in adolescents and young adults: A study of parental reports.” *PloS one* 13.8 (2018).

"hypothesis" function). Then, the model calculates the "loss," or the difference between its guess against the actual word from the training data. Finally, the model adjusts its guess very slightly, minimizing the difference between the guessed word and the actual word. Though the adjustment it makes is miniscule, with enough time and computational power, the model will make enough adjustments to eliminate the difference between the guess and actual word. It will move toward generating a prediction that is increasingly plausible.

This paper does a deep, close reading of the statistical methods underlying this prediction process, which results in a kind of approximation of language, and connects this approximation to what Trans Studies scholars define as a central desire to *pass*. Andrea Long Chu claims that Trans Studies "requires that we understand—as we never have before—what it means to be attached to a norm, by desire, by habit, by survival."[3] Could AI-generated text, as an approximation, or even normalization, of its training data, be used to study the attachments to norms that characterize the fear of transness?

This paper takes the attachments and investment in norms, which has traditionally been a tool of oppression, as a creative resource for, in Sedgwick's words, "extracting sustenance from the objects of a culture… whose avowed desire has often been not to sustain."[4]

[3] Chu, Andrea Long, and Emmett Harsin Drager. "After trans studies." *Transgender Studies Quarterly* 6.1 (2019): 108.

[4] Sedgwick 35.