Contents

1 trans studies and the materialization of gender

1.1 TODO total rewrite

How trans materialism affects our approach toward gender representation.
overview:

☒ opening: disagreements globally about what is gender, affecting transgender rights movement.

- Both pro and anti trans turn on the fluidity of gender.

- thesis: bringing trans materialist theory to bear on the study of gender bias – that gender is a social and behavioral phenomenon, not an internal identity. How does this affect the labels and contexts we give to gender?

- defining gender:

  ☒ Butler: identity/self only comes into being through action, but it is a result, a construction, of a psychological process, through a series of disavowals.

  ☒ divergence model

- □ amin: we ought to treat gender as a behavioral/social phenomenon,
  - ∗ otherwise we keep kicking the can of gender nonconformity down the road.
  - ∗ we need recognition in particular for feminized masculinity
- – nonbinary identity is unethical,
  - ∗ because it makes visible already mainstreamed kinds of gender
  - ∗ because it severs gender from sexuality and from sex
  - ∗ leading to today's anti-trans discourse

- word embeddings, underrepresentation:
  - van Loon: freqency of terms has effect on positivity
  - what is not readily marked is assumed to be invisible or inoperative
    - ∗ critique of Charlesworth: the ways that class is being theorized in trans studies and how that affects intersectionality.

- counterfactual augmentation methods
  - giving the same contexts to male, female, nonbinary.

- is it possible to have an unmarked gender that is not dominant?
  - the problem is that we do not denaturalize the sexes
    - ∗ Wolfe, Caliskan: males with masculinity, females with femininity.
    - ∗ Halberstam: there is female masculinity.

- it's about switching identity for presentation.
  - all of them should have the same contexts? male, female, nonbinary? But we should have specificity on presentation. That someone of a more feminized gender might have typically feminine traits like longer hair?

How do you develop a categorization schema based on social relations and behavior?

(we need to talk about the harms that gender does when it is visible versus when it is not visible. Hunter Shaffer saying that they will deal with the M on their passport is a statement of privilege, as a white woman, she can say that. But what about the trans migrant?)

1.2 reading notes

Wolfe and Caliskan

Guo and Caliskan

Shiva Omrani Sabbaghi, Robert Wolfe, and Aylin Caliskan. 2023. Evaluating Biased Attitude Associations of Language Models in an Intersectional Context. In Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES '23), August 29, 2023. Association for Computing Machinery, New York, NY, USA, 542–553. `https://doi.org/10.1145/3600211.3604666`

Devinney et al. 2024:

- goals: studying bias about power symmetries in models that have been optimized or finetuned to be safe and unbiased.
- methods: the interesting thing is that they don't use the regular metrics to measure bias, they use close reading, the EQUITABL method.
- findings: that llms are reluctant to discuss identity, which is harmful; that unmarked dominant identities are associated with positivity more than marked/marginal ones. –> unmarked dominant identities are more associated with positivity than marked/marginal ones.

1.3 introduction

This paper brings theorizing from the disciplinary field of Trans Studies to bear on the study of gender bias in NLP. It considers conversations from Trans Studies about defining gender as an external presentation or expression, which contradicts progressive, trans-affirming discourse that emphasize gender identity.

From within this framework, this paper considers the implications of marking gender as identity or as expression. It then reviews some recent work in evaluating gender bias that makes some progress in this area, and makes suggestions for future research.

1.4 popular disagreements on gender

There is a global urgency to understand and conceptualize gender, how it ought to be defined and legislated. In various countries around the world, gender has become a national issue, and a fear of so-called "gender ideology" has served as a rallying cry for far right groups to garner support [Butler 2023].

For example, in the European continent, far right groups have used transphobia to prop up other fears, such as those against migrants, in what some argue is a white supremacist campaign [Indelicato and Lopes, 2024]. In the African continent, the same fear has been deployed in anti-colonial discourses, asserting that gender ideology is a dangerous import from the West that threatens traditional gender roles and the nuclear family [Butler 2023].[1] The past few years have seen a wave of anti-LGBT+ legislation, for example the law in Uganda that penalizes "aggravated homosexuality" with the death penalty [BBC 2023].

In North America, where I am located, a mounting wave of transphobia has recently propelled a series of executive orders with titles like "Defending Women From Gender Ideology Extremism And Restoring Biological Truth To The Federal Government" and "Keeping Men Out of Women's Sports. These initiatives double down on the primacy of sex over defining sex as "binary and biological," and relegating gender to a personal feeling that has no bearing on sex identification [The White House 2025a, The White House 2025b].

The push to solidify sex as "binary and biological" is a reaction to the fluidity, ~~nebulousness, and slipperyness~~ of how gender is conceptualized in trans affirming discourse. The American Psychiatric Association defines gender identity as "a person's inner sense of being a girl/woman, boy/man, some combination of both, or something else" ["What is Gender Dysphoria?" 2025]. Similarly, the World Health Organization defines gender identity as "a person's innate, deeply felt internal and individual experience of gender," and constrasts it to biolgoical sex, adding that gender identity "may or may not correspond to the person's physiology or designated sex at birth" ["What is Gender Dysphoria?" 2025, "Gender and health" 2025].

In this view, gender is associated with gender identity, an internal phenomenon. Sex, by contrast, is a biological fact, one that can be altered to reflect a person's internal sense of gender. This view inspires transphobic reasoning around the issue of medical care–if gender is an internal phenomenon, then why do people feel the need to physically transition? In other words, if gender identity can actually diverge from external sex characteristics (genitals, facial hair, etc), then sex has no bearing on gender identity, which means that transgender people shouldn't need to transition at all.

This paper argues that one particular difficulty of mitigating gender bias

[1] Ironically, many of these anti-gender ideology movements are funded by Western religious groups, particulary the Vatican and Evangenlical organizations in the USA [Butler 2023].

in NLP has to do with the general disagreement about what gender is in the first place.

This paper considers current conversations in Trans Studies that pursue a conceptualization of gender that pushes against this ambiguity. It then explores how these conceptualizations might be applied to current methods in NLP.

1.5 On gender categories

I will briefly trace the definintion of gender over time, taking into account how it has been conceptualized in relation to closely associated notions of sex and sexuality.

1.5.1 performativity

Contemporary liberal notions of gender descend from a widely accepted view that gender is a social construction rather than biological reality. This view can be traced to Judith Butler's seminal theory of gender performativity, which dissolves the traditionally strict confines and causal relationship between sex, gender, and sexuality. Butler's theory of gender performativity stipulates that gender is not, as widely assumed, based on a biological reality, and deterministic of a sexuality. Rather, Butler argues, gender is an *effect* of social norms that govern behavior. According to this theory, gender is made real through its expression, an expression that is compelled by social expectations according to each gender role.

35 years after Butler's theory, the understanding of gender as a kind of performance or behavior (one that, Butler is careful to point out, is not at will, but deeply compelled by social rules) has developed into a diverged model that separates identity from presentation. In this model, gender identity is an internal sense or feeling of identification with male, female, a combination of the two, or nonbinary. Additionally, gender identity is separate from gender expression, which denotes the masculine, feminine, or androgynous traits that are visibly apparent and enacted. Finally, sex is understood to be a bodily component, though it cannot be said to be based firmly in a binary, nor does it necessarily prioritize any biological marker, like the presence of external genitals, sex hormones or chromosomes.

1.5.2 categorical splits

The development of the gender identification has, according to Queer and Trans Studies scholar Kadji Amin, created a domino effect where gender

is increasingly idealized and divorced from bodily expression. Amin argues that identity terms often used within transgender discourse, particularly the terms "cisgender" and "nonbinary," reflect an idealized conceptualizations that actually re-stigmatize non-normative gender expressions.

This idealization has to do with a split between gender and sexuality, which ~~Amin, Butler, and other~~ Queer and Trans Studies scholars argue cannnot be neatly separated. Amin, who opts for the hyphenated term, "gender-sexuality" to indicate the imbrication of gender expression and presentation with sexual orientation, argues that this "divergence model" between gender and sexuality is indicative of another series of splits between categories like heterosexual and homosexual, and cisgender and transgender. According to Amin, while homosexuality used to incorporate gender-variant forms of sexuality, for example, male femininity in the figure of the "queen" or the "fairy," and female masculinity in the form of the "butch" or the "stud," the emergence of transgender offered a new category for "all manner of gender-bending" [Amin].

Another significant effect of the categories is that they create idealized forms of normativity that do not accurately describe any gender-sexuality in lived reality. This is because normative categories like "heterosexual" and "cisgender" form as a reaction to the emergence of non-normative identities. Historiographers of sexuality generally agree that the concept of heterosexuality arose as a reaction to the definition of homosexuality in the 19th century by sexologists [Foucault 1978, Halperin 2002].[2] Just as heterosexuality emerged "belatedly, as a normative ballast against homosexuality," Amin argues, so did the concept of cisgender, which responds to the coinage of the term "trangender" in the 1990s [Amin 110-11].[3] As a result, rather than descriptive categories, the terms heterosexual, cisgender, and binary represent fictions, "increasingly idealized and uninhabitable normative categories," created to balance the non-normative identities of homosexual, transgender, and nonbinary [Amin 116].

[2]Historiographers of queer history generally agree that sexual orientation is a modern phenomenon, developed by sexologists in the 19th century. See Richard von Krafft-Ebing, *Psychopathia sexualis* (1886); Havelock Ellis, *Studies in the Psychology of Sex.* Vol. 2 Sexual Inversion. 1900.

[3]The terms "homosexuality" and "transgender" here are treated as terms rather than concepts; in reality, homosexuality exists since classical times, depicted particularly in Egyptian, Greek, Hindu, Buddhist cultures. Despite this representation, the identity of "homosexual" only emerged in the 19th century as a result of sexological research in Europe.

1.5.3 cis and binary are idealized and uninhabitable

As a reaction to non-normative categories, the normative categories of "heterosexual," "cisgender," and "binary" imply idealized notions of gender-sexuality that are difficult to inhabit because there is not inner feeling of identification. This lack of identificatory feeling explains why, for example, many apparently cisgender people find it difficult to identify with the "cisgender" and "binary" nominatives. While it is relatively simple to name the experience of discomfort in one's embodied gender-sexuality as gender dysphoria, it is less straightforward to name and describe the experience of what is unremarkable, routine, and even banal. There is no salient feeling that describe the experience of being comfortable in one's gender-sexuality.

1.5.4 nonbinary is contrasted with sex

One effect of idealizing gender-sexuality is place it as a psychological identification. On this point, at least, both the transphobic and the trans-affirming perspective align. For example, a recent executive order from the United States White House unambiguously contrasts "sex" from "gender identity," which it defines as:

> A fully internal and subjective sense of self, disconnected from biological reality and sex and existing on an infinite continuum, that does not provide a meaningful basis for identification and cannot be recognized as a replacement for sex. ["Protecting women" 2025]

Similarly, the World Health Organization (WHO), who has codified gender dysphoria as "gender incongruence," defends the right gender-affirming care ["Gender incongruence and transgender health in the ICD"]. The WHO affirms that sex is separate from gender, which is a "socia construct," and that gender itself is separate from gender identity:

> Gender and sex are related to but different from gender identity. Gender identity refers to a person's deeply felt, internal and individual experience of gender, which may or may not correspond to the person's physiology or designated sex at birth. ["Gender and health" 2025]

Both those who deny and affirm the rights of transgender people agree on one thing: that gender identity is distinct from expressed and embodied forms of gender.

1.5.5 feminized men

As it increasingly diverges in categories, Amin argues, gender identity becomes buried "deeper and deeper within the private recesses of the self, where it increasingly disavows any relation to the social" [Amin 116]. In other words, idealized categories like cisgender and nonbinary make non-normative gender identities less and less visible. For example, in the identity of the "femme" (feminine) presenting nonbinary person describes a woman assigned with a feminine gender expression that nonetheless identifies as nonbinary.

The effect of prioritizing non-normative gender as an *identity* that is invisible, that is not immediately apparent despite being marked, is to marginalize gender as *expression*, which re-stigmatizes those gender groups that already experience social stigma. Amin points to feminized genders to make this point–the queen and the fairy, for instance. Defining gender as an external expression or embodiment is more likely to make visible and validate those gender which experience stigamtization based precisely on how those genders present outwardly.

1.6 default is not marked

I ended the previous section by saying that, according to Trans Studies, gender as internal concept of identity contrasts to gender as expression, presentation, or behavior.

The difference has to do with a kind of stigmatization that accompanies non-normative genders that are externally expressed. Amin gives the example in homosexual culture, how normative masculine or "butch" gay men have higher social status than feminized gay men, due to the cultural capital of masculinity. The ways that these nomatives work, Amin explains, is by taking what is most visibly palatable or normative within each category.

There is a direct relationship between what Amin describes as presentation, expression, or embodiment in gender and gender's representation in language in NLP.

Existing evaluations of LLMs show that the the default subject category is the straight white male, and that other subject categories, like woman, black, and homosexual, are less prevalent in both training data and model outputs [Devinney, Guo and Caliskan]. Additionally, research on word-vector evaluation methods shows, for example, that the relative frequency of a particular identity group in the training data will have an effect of that group's association with positive and negative terms [van Loon].

So it seems that presence in the training data has a positive effect on whether subjects are represented in a positive way. In other words, there is a direct correlation between a subject's visibility and their cultural value.

We can reasonably assume that making marginalized subjects more visible in the training data, to be on par with the frequency of representation of dominant subjects, will be sufficient for mitigating bias about these groups in model outputs.

But, in order to do so, gender ought to be conceptualized as an outward, expressive phenomenon, rather than an identity category.

In other words, adding additional labels for "masculine" and "feminine", such as "masc" and "femme," rather than just "man," "woman," or "male" and "female." We might have targets and attributes that describe gender as physical and behavioral, rather than as internal or psychological.

The problem with the univeral, unmarked subject is precisely that – they are not marked. The unmarked subject is not marked yet is visible and dominant. The over-representation of straight white men is not due to a persistant label of "heterosexual white man" that appends each and every predication. Rather, it's due to a proliferation of predication that already surround him, that define him. The white man does not a descriptive label as such, he simply is.

1.7 some suggestions

Target-attribute ought to focus on gender as presentation rather than identity.

In "Fighting Bias with Bias," Reif and Schwartz [2023] demonstrate a promising approach: amplifying rather than reducing bias in a model's training dataset. They point out that bias reduction techniques are not very effective, that "filtering can obscure the true capabilities of models to overcome biases, which might never be removed in full from the dataset" [Reif and Schwartz 2023]. Instead of reducing, they follow the work of Stanovsky et al. [2019], who intensify bias by including phrases like "the pretty doctor" so that a model will interpret "doctor" to be female.

1.8 further study

Using qualitative methods to close-read gender attributes.

Devinney et al 2020 combine quantitative and qualitative methods, distant and close reading, to study gender and racial bias in LLMs. The methodology, EQUITBL (Explore, Query, & Understand Implicit Textual Bias in

Language data) developed by Hannah Devinney and Henrik Björklund, uses topic modeling and data visualization to "produce a view of the data of a size suitable for qualitative analysis" (Devinney et al 2020).

1.9 conclusion

There's a central contradiction in the theorizing of gender – as something that is a social and behaviorial phenomenon (ala Butler) and at the same time an identity and expression of identity. (We see this problematized in the right as a confusion about the desire for gender affirming surgery if gender is really not already determined by sex.)

"What is socially relevant is transition—-a shift in social gender categories, whatever they may be—-not identification—-a personal, felt, and thereby highly phantasmic and labile relation to these categories. Identification is the psychic process that makes the interval between the individual and the social apparent; it is not the site of their suture." (Amin 115).

1.10 bank

Trans liberalism is a problem because it overlooks class.