Contents

1 some myths about bias: a queer studies reading of gender bias in NLP

1.1 abstract

This paper critically examines gender bias in large language models (LLMs) through a critique of binary forms, adopted from the field of Queer Studies. It argues that many existing bias detection and mitigation techniques in Natural Language Processing (NLP), particularly those that use word vectors and gender-swapping, impose binaristic forms of thinking about bias that go beyond outdated assumptions of gender as a symmetrical and stable form. Drawing from Queer Studies, the paper highlights two "myths" about gender bias: first, that bias is categorical, taking on a dichotomous variable ("yes/no," "pass/fail"), and second, that it is zero-sum, that discrimination can be levelled to have a neutral effect between groups. Due to their operationalization of binarisitic thinking, each of these myths effectively reduce and flatten bias into a measure that fails to reflect the workings of semantics, discrimination, and prejudice in language. The paper concludes by suggesting that bias mitigation in NLP should focus on diversifying gender expressions, rather than attempting to neutralize or equalize them. By considering humanistic critiques of the binary, NLP may fashion more inclusive and intersectional approaches to mitigating bias in language systems.

CCS CONCEPTS • Computing methodologies → Natural language processing; • General and reference → Metrics; Evaluation; • Applied computing → Arts and humanities.

Additional Keywords and Phrases: natural language processing, gender bias, queer studies, bias evaluation, bias mitigation

1.2 introduction

This paper analyzes methods for evaluating and mitigating gender bias in Large Language Models by drawing from current conceptualizations of gender from the humanities. It argues that mitigating gender bias requires understanding gender as an identity and operation that has been vigorously theorized in fields that specialize in sex, gender, and sexuality, like Women's Studies, Gender Studies, Feminist Studies, Queer Studies, and Trans Studies. It incorporates domain-specific knowledge from the field of Queer Studies in particular to analyze how embedded assumptions about gender binaries drive current bias evaluation and mitigation methods.

[why QS]

I have chosen the field of Queer Studies as the foundation for my critique because this field offers a socio-constructivist model of gender that emphasizes critical analysis of the binary as an ideological structure. This model, Judith Butler's theory of gender performativity, which inaugurated the field in the early 1990s, influences the perception today that gender is a social and behavioral phenomenon, rather than a biological reality [Butler 1992]. Since Butler's theory, the distinction between gender as a social operation and sex as a physical embodiment, and the subsequent dissolution of a binary model of gender difference, have been validated in biology, neuroscience, and psychology [Ainsworth 2015, Hyde et al. 2019, Joel 2020]. At the same time, Queer Studies has continued to debate and problematize the sex/gender division, with disagreement about how this division affects the visibility of Transgender subjects in particular [Amin 2022]. Additionally, discussions of sex, gender, and sexuality have expanded to include race, class, (dis)ability, and other aspects of identity, and have risen to prominance in Intersectional Feminism, which is now accepted as a standard paradigm for critical analysis in the humanities [hooks 2000, Munoz 2009, Klein and D'Ignazio 2024].

[motivation]

This paper adopts a humanistic method of problematizing implicit assumptions in research to frame a critique of methodologies used to study and mitigate bias in NLP. It speaks to recent work in NLP and ML that examines implicit assumptions in research [Klein and D'Ignazio 2024, Birhane et al. 2022, Devinney et al. 2022]. Birhane et al 2022, for example, challenge assumptions of the ML field as "value-neutral," and adopt a close reading of ML papers to annotate these values. They find that papers encode values

implicitly in aspects like project choice, justification, and researcher affiliations, which determine what kind of research gets done and what groups will benefit [Birhane et al. 2022]. Taking this perspective to the study of bias, Blodgett et al. 2020 study how NLP papers operationalize and analyze the concept of social bias (including racial and gender bias), finding that most of them do not define what bias is, how it is harmful, or explain why reducing bias is important, for what groups. To resist "self-evident statements of 'racial bias'", which are always underpinned by implicit values, Blodgett recommends consulting interdisciplinary work on bias, particularly work that addresses how language and social discrimination are coproduced, and the role that language plays in creating and maintaining social hierarchies.

[urgency]

This work furthers an area of NLP research that is already robust with critiques of bias detection and mitigation techniques. While many studies have pointed out how such methods are ineffective or counterproductive [Gonen and Goldberg 2019; Blodgett et al. 2021], which others have attributed to a misunderstanding of how gender bias operates in language [Devinney et al. 2022, Hitti et al. 2019, Nemani et al. 2023, Meade et al. 2022, Caliskan et al. 2022], none have, to my knowledge, explored their ineffectiveness by critiquing the binary as a conceptual model.[1] Those that do mention binaries, understandably do so in the context of gender binary, i.e., male/female [Hitti et al. 2019, Nemani et al. 2023].

This paper, however, considers the binary as not just a way of categorizing gender identity, but as a deeper structure of thought. Borrowing from the insights of Queer Studies, this paper considers how the binary, in organizing information into a dichotomous model (yes/no, male/female), determines the relationship between terms. As Queer Studies scholars like Judith Butler, Even Kosofsky Sedgwick, Jack Halberstam, and Kadji Amin argue, the binary works by positioning its terms into a symmetrical and oppositional relationship, a relationship that imposes a dynamic of contrast, hides underlying power relations, as well as delimits what can be represented against that which is unrepresentable [Butler, Sedgwick, Halberstam, Amin]. This paper argues that the binary, as a form of thinking that encodes power relations between terms (and between what is excluded from them), implicitly structures the conceptualization of bias in NLP into two "myths" (1) that bias is categorical, and (2) that bias is zero-sum. These myths reveal that some of the goals in evaluating and mitigating bias, such that equity is the

[1] Lauren Klein and Catherine D'Ignazio in "Data Feminism for AI" is the call to "rethink binaries" [Klein and D'Ignazio 2024]

same as equality, are based on theorizations of the binary as a symmetrical and stable form.

[overview]

In what follows, I review current literature on gender bias in NLP, outlining different conceptualizations of how bias appears in language. Then, from Queer Studies, I review the critical analysis of binary models of organization, and how they necessitate certain exclusions that inevitably emerge to disrupt the apparent stability of the binary. Subsequently, in the main section of the paper, I apply this critique to a reading of bias evaluation and mitigation techniques that center on word vector technology like WEAT (The Word Embedding Association Test) [Caliskan et al. 2017], and DeBias [Bolukbasi et al. 2016], as well as those that use gender swapping and Counterfactual Evaluation [Zhao et al. 2018, Meade et al. 2022, Nemani et al. 2023]. Finally, I close by pointing to some promising work in current NLP that expands beyond the limitations of a binary model and operationalize that model in capacious and productive ways.

1.3 literature reviews

1.3.1 existing schemas of gender bias in NLP

Existing research defines bias by how it is expressed in language and by its social effects. Hitti et al. [2019], who examine how bias expresses in language, divide bias into structural and contextual types. Structural bias concerns bias that results from grammatical structures, such as pronouns that assume a male antecedent ("A programmer must always carry his laptop with him"), while contextual bias concerns bias that results from social and behavioral stereotypes ("Senators need their wives to support them throughout their campaign") [Hitti et al. 2019]. Moving from language expression to social effects, Nemani et al. [2023] classify bias by the particular kind the implication it has for a specific social group, and organizing bias into the categories: "Denigration," "Stereotyping," and "Under-representation." Denigration refers to the use of derogatory language such as slurs; stereotyping refers to prejudice about a particular social group; and under-representation refers to the relative dearth of information about a particular social group [Nemani et al. 2023]. Similarly, Barocas et. al [2017] divide bias into "allocative harms," where resources are withheld from certain groups, and "representational harms," where certain groups are under-represented or stereotyped.

This paper focuses on the social effects of bias, and adopts Nemani et

al.'s useful tripartite scheme for organizing bias. As I demonstrate below, bias often exceeds a categorical measure, so that having multiple categories like "denigration," "stereotype," and "representation" will yield more precise and illustrative analysis. Additionally, current work on bias which does not distinguish between these categories tends to conflate one with another, so that, for example, stereotype is considered equivalent to denigration. These tendencies, which I argue are a result of binary thinking, collapse different types of bias into one totalizing frame. For example, the common assumption that all bias is negative and harmful will likely categorize the association between association between women and terms like "mother" as denigration, without considering the roles of stereotype and under-representation. Such conflations lead to mitigation strategies that are less specific, tailored to the particular type of bias, and therefore less effective.

1.3.2 queer studies on binaries

While bias detection and mitigation methods in NLP aim for an elimination of bias, Queer Studies field has problematized the idea that inequality can be eliminated from social systems.[2]

One central concern for Queer Studies is the problematization of the gender binary, and of binary structures generally, which can be traced to Judith Butler's theory of gender performativity, famously outlined in her first book, Gender Trouble: Feminism and the Subversion of Identity [1990], but more robustly theorized in her follow up work, Bodies That Matter: On the Discursive Limits of Sex [1993]. Butler's theory of gender performativity stipulates that gender is not, as widely assumed, an inner truth or biological reality. Rather, it is an ideological construction constituted by societal norms that manifests in behaviors. According to this theory, gender is created or made real through its expression in gender roles.

Despite the popularity of Butler's theory, which some researchers in NLP have used to explain the constructed nature gender [Devinney et al. 2022], a crucial detail of her argument goes relatively unnoticed. This detail is that gender, for Butler, is not merely an effect of social conditioning. Rather, it is form of social regulation, a power structure that that effectively partitions social roles with the effect of "domesticat[ing]... difference" within a

[2]In Queer Studies, there are two general approaches for proceeding under these conditions. First: to create strategies of thriving within unjust dynamics, finding alternative modalities of survival, liberation, and joy: See Butler [1993] and Munoz [2009]; Second, to explore and outline the contours of stigmatization, shame, and oppression from within those less palatable spaces of inequality: see Edelman [2004] and Love [2009].

hierarchical social order [Butler 1993].

As many Queer Studies scholars point out, one way that social hierarchies are reinforced is through the imposition of categories such as binaries, for example, "male/female," and "heterosexual/homosexual." Binaries create an apparent stability through delineating two entities (such as “male” and “female”) into an ordered relation. One effect, according to Queer Studies scholar Eve Kosofsky Sedgwick [1990], is to bring its terms into legibility through contrast and opposition. As Sedgwick explains, in the binary "heterosexual/homosexual", the term "heterosexual" is not simply symmetrical to "homosexual," but rather, depends on "homosexual" for its meaning through "simultaneous subsumption and exclusion." In fact, as Kadji Amin and historians of sexuality assert, the concept of a heterosexual identity only emerged as the definition of homosexuality was being established by sexologists and psychiatrists in the late 19th and early 20th centuries; heterosexuality, in other words, appears on the scene for the purpose of outlining the limits of what was then taken to be a perverse and abberant orientation, "as a normative ballast against homosexuality" [Amin 2022].[3] In other words, one term, such as "heterosexual," achieves its definition by circumscribing the content of the other term in the binary.

The meaning of the binary terms are achieved by the dynamics between what is represented and what is excluded. Butler refers to this as the dynamic between the binary and its "necessary outside," an element that is excluded from the binary, whose exclusion enables the binary's operation. For example, in the "heterosexual/homosexual," not only is "heterosexual" defined in contrast to homosexual, but "homosexual" itself is defined against a sexuality that is not representable from within that schema. The binary gains its definition precisely by what is excluded, what Butler describes as "a domain of unthinkable, abject, unlivable bodies" [1993], from that conceptual system.[4]

In this reading, then, binaries attempt to stabilize and delimit concepts into relations of contrast and opposition. Although they work by illustrating a certain symmetry, the terms of the binary are not symmetrically related. Their apparent stability always masks an underlying imbalance. However, despite their constraining nature, binaries, in Sedgwick's words, remain "peculiarly densely charged with lasting potentials for powerful manipulation"

[3] Citations to Jonathan Ned Katz and David Halperin.

[4] Scholars in fields like Black Feminist Studies and Trans Studies explore how such exclusions operationalize race for the creation of gender orders. For example... Spillers and Snorton.

like Riley C. Snorton has used a similar argument for race, arguing that

– a topic I will return to in this paper's discussion [1990].

In the next section, I argue that these assumptions about stability, symmetry, and equal relations influence how NLP has conceptualized gender bias. I explore two myths of gender bias: (1) that bias is categorical, and (2), that bias is zero-sum.

1.4 myth 1: bias is categorical

Binary thinking affects the study of gender bias in NLP: it rallies different kinds of bias into a categorical measure for detecting and evaluating bias. This kind of thinking is apparent in some bias mitigation and evaluation techniques that leverage word vectors, such as WEAT (The Word-Embedding Association Test), which has influenced subsequent vector-based methods like SEAT (Sentence-Embedding Association Test) and FISE (Flexible Intersectional Stereotype Extraction procedure) [Caliskan et al. 2017, May et. al 2019, Charlesworth et. al 2024].

The WEAT metric's development, and particularly the way it adopts concepts from across disciplinary understandings, illustrates a reductive conceptualization bias that limiting the kinds of results bias evaluation and mitigation techniques can achieve.

First, the concept of "bias" is adopted from a machine learning context to study social phenomena. In machine learning, bias is a single measure that captures the accuracy and correctness of model output, and it is measured by subtracting the true value of an output from its expected value. Bringing this concept to a social context, the WEAT authors assert that, "In AI and machine learning, bias refers generally to prior information, a necessary prerequisite for intelligent action. Yet bias can be problematic where such information is derived from aspects of human culture known to lead to harmful behavior" [Caliskan et al. 2017]. In WEAT, this understanding of bias as "prior information" which affects accuracy is summarily transferred into a measure of that can "lead to harmful behavior" [Caliskan et. al 2017]. This move assumes that bias is equivalent to one particular type of bias, to denigration, which at best, ignores other types of bias, like stereotypes and under-representation, and at worst, collpases these, which are highly variable on positionality and context, into a single score.

Second, in another transaction between disciplines, WEAT takes a concept from social psychology into to vector space. In social psychology, the Implicit Association Test (IAT) [Greenwald et al. 1998] measures the association that a test subject makes between a particular identity group and an evaluative term, like "good" or "bad." Here, the subject will categorize pho-

tos of people with one of two labels, such as "fat" or "thin," using their right or left hands which contain a response key for the label [Greenwald et al. 2011]. In the next round of the test, they will be shown different words and categorize those words as "good" or "bad," again using the right or left hand to press a response key that indicates the category. The test will then proceed with two more rounds with similar prompts, but the response key will shift between hands. The test assumes that the response time for selecting a response key like "fat," correlates with the evaluative term, such as "good" or "bad," that had just corresponded to that response key in the previous round. The test developers conclude that, "one has an implicit preference for thin people relative to fat people if they are faster to categorize words when Thin People and Good share a response key and Fat People and Bad share a response key, relative to the reverse" [Greenwald et al. 2011].[5]

In applying IAT to vector space, WEAT uses co-sine similarity as a correlative to response time, so that a shorter distance between vectors inidicates an implicit preference and a longer distance indicates an implict aversion. The results of the test, the authors point out, bear out all of the experiments done with the AIT. They show, for example, "stereotype biases" such as "ingroup/outgroup identity formation," as well as "the gender distribution of occupations." These results, the authors argue, suggest that bias may be "a simple outcome of unthinking reproduction of statistical regularities absorbed with language" [Caliskan et al. 2017].

Although the WEAT is useful for measuring associations between terms, its view of bias as a categorical value, which it inherits from its progenitor the IAT, limits its ability to capture different kinds of bias and how they may influence one another. The AIT's approach toward bias as something that can be represented as categorical value, as present or absent, effectively imposes an evaluative measure on top of a detection one. It is a small step from reporting that a person has a stronger association with a certain identity group over another, to claiming that the association measures social bias. However, to the extent that an association can be detected does nothing to reveal the harmfulness of that association, not to mention its particular quality or effect–having to do with stereotype, representation, or denigration, for example.

This subtle imposition of evaluation on detection, as a result, fundamentally misses the ways that bias is conceptualized and operationalized in

[5]The test is not without its critiques within the field of Social Psychology, for example that it lacks "construct validity," that results vary widely and it has no effect on explicit attitudes. See Schimmack [2021] and Karpinski [2001].

language, which is revealed by downstream effects of using WEAT-based methods. One example shows how the bias as under-representation becomes conflated with that of denigration, to the confusion of researchers. For example, in a study using word vectors, names that are over-represented exhibit a higher positivity score, while those that appear fewer times show a negative score [Wolfe and Caliskan 2021]. Here, the under-representation of certain group names, those of typically minority groups, has a detrogatory effect on their portrayal, thus perpetuating their marginalization. To correct for this effect, a subsequent study [van Loon et al. 2022] controls for the variable of term frequency, augmenting the number of times minority names are mentioned in the training data. The authors note that the solution is "unintuitive", cautioning that, "if other biases we don't know about are also introduced by the use of word embeddings, we might not be able to rely on standard sociodemographic controls to fully address them [van Loon et al. 2022].

1.5 myth 2: bias is zero-sum

Rallying all of bias into y/n not only obscures the differences between the types of bias, it also suggests that bias is a quality that can be extracted and separated from text. It is a short step from the perceiving an aspect as present or absent, to believing that it can be excised.

The this excisable quality emerges in another word vector-based technology, "DeBias," which is a mitigation strategy that attempts to deduct or neutralize bias from vector space. Developed by Bolukbasi et al. [2016], the method works by calculating "gender subspace" or "gender direction" for certain word vectors that have gender connotations. Words such as "gal", "guy," "programmer," and "babysitter," have associations to a certain gender. Then, depending on whether the terms are gender specific or gender neutral ("gal/guy" is gender specific, while "programmer" and "babysitter" are gender neutral), the terms are either "equalized" or "neutralized." Terms that are neutralized have values closer to zero in the gender subspace, while terms in the equality set are made equidistant from the gender neutral terms. The goal is that gender neutral terms are not more associated with one gender over another, but remain equally associated with both genders. For instance, the developers explain that, "after equalization babysit would be equi-distant to grandmother and grandfather and also equi-distant to gal and guy, but presumably closer to the grandparents and further from the gal and guy" [Bolukbasi et al. 2016].

Criticism of this DeBias shows, however, that gender bias cannot be ex-

tracted from a text like a single thread from a cloth. Gonen and Goldberg [2019] in particular claim that the results are "superficial," arguing that, "While the bias is indeed substantially reduced according to the provided bias definition, the actual effect is mostly hiding the bias, not removing it. The gender bias information is still reflected in the distances between 'gender-neutralized' words in the debiased embeddings, and can be recovered from them" [Gonen and Goldberg 2019]. For example, they find that after DeBiasing, words like "nurse," while no longer associated with "explicitly marked feminine words," maintains its proximity to "socially-marked feminine words," like "receptionist," "caregiver," and "teacher" [Gonen and Goldberg 2019].

Similar to WEAT, this method approaches bias as an absolute quality, collapsing examples of stereotype with denigration. A closer attention to the particular type of bias would help to explain which kinds of associations are actually undesirable. For example, the terms "math" and "delicate," as Gonen and Goldberg explain, "have strong stereotypical gender associations, which reflect on, and are reflected by, neighbouring words" [2019]. But while these terms carry stereotypical associations, and stereotypes are reductive, they are not in themselves harmful. The harm comes from making further associations between these terms, for example, if the association with delicate also leads to one of inferiority or weakness. Such an association marks the feminine gender as with a need for protection, which is a starting place for patriarchial associations that are controlling and/or belittling in the name of protection.

The idea that gendered terms can operate "neutrally" or "equally" across contexts influences other bias mitigation techniques which are based in gender swapping. Counterfactual Evaluation and Winobias, for example, measure gender bias by swapping gender terms such as pronouns and tests their associations with particular attributes and its effect on on model performance [Nemani et al. 2023] [Zhao et al. 2018]. Because the results of these assessments reflect only a change in gender, it is reasonable to assume that they may be used to measure gender bias. However, these methods do not take into account how gendered terms carry connotations that do not make them equivalent or able to be substituted one for the other. For example, Devinney et al. [2022] explain that in the word pair "bachelor" and "spinster," the term "spinster is pejorative while bachelor is not," pointing out that "there is no such thing as a spinster's degree." Many such terms carry with them denigrating associations of gender that are perpetuated into word vector space, so that any gender swapping techniques will implicitly carry these associations along with the change in gender.

These methods have in common the assumption that gender is a zero-sum phenomenon. They take this assumption from the binary form, that because feminine are diametrically opposed and symmetrical. Therefore, it is only a question of equalizing stereotypical associations between masculine and feminine words. In reality, however, the relation between gendered terms is not symmetrical: associations may be simply stereotypical or more directly denigrating, or they may lead to other terms that carry these associations. Treating all gendered terms as symmetrical overlooks the complex and perhaps untraceable ways that bias operates across embedding space.

In the next section, I explore possibilities for working within these constraints.

1.6 discussion

This paper has shown some ways that the binary thinking inspires assumptions about bias–that it is categorical, a single measure that collapses "prior information" with types of bias like stereotype and denigration, and that it is zero-sum, where gendered words can be equalized or neutralized in terms of valence and connotation.

In what follows, I will point out two areas of promising research.

There are methods in NLP that reformulate the traditional binary to mitigate gender bias. These methods, what Devinney et al. [2022] call "trans-inclusive methodologies," expand the traditional binary. For example, Hansson et al. [2021] incorporate a gender neutral pronoun "hen" in Swedish into their Wino-gender dataset. Additionally, Dinan et al. [2020] expand the classification of gender in their dataset to include "neutral" and "unknown." Crowd-sourced and participatory datasets also contribute to this effort, namely when they are done by participants of the community, like WinoQueer [Felkner et al. 2023]. Such work take exploratory and crucial steps on the path to gender equity in language systems.

But this paper does not recommend that we leave the binary behind. Binaries remain, in Sedgwick's words, "peculiarly densely charged with lasting potentials for powerful manipulation" [Sedgwick 1990]. While the dimoprhic structure of the binary imposes symmetry through opposition, it also opens the possibility of combinations through relay. Jack Halberstam elaborates on this point, explaining that "Gender's very flexibility and seeming fluidity is precisely what allows dimorphic gender to hold sway" [1998]. Because there are two poles of gender, so to speak, then gender expressions can vacillate between the two, what Halberstam describes as "multiply relayed through a solidly binary system."

Some examples of recent work in NLP that explores how this potential by pushes the strict confines of the binary model. One strategy harnesses stereotypes to its advantange, to amplify (rather than reduce) bias in a model's training dataset. In "Fighting Bias with Bias," Reif and Schwartz [2023], following the work of Stanovsky et al. [2019], intensify bias by including phrases like "the pretty doctor" in the training data. The idea is that a phrase which mixes stereotypes, such as feminine traits ("pretty") with masculine occuptations ("doctor"), will result in gendering "doctor" as female (or alternatively, describing a male gender as "pretty", which also disrupts stereotype) [Stanovsky et al. 2019]. According to the researchers, bias amplification succeeds where attempts of reduction have failed due to the capacity of language models to generalize from biased over "unbiased" examples: "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].

Such projects resist assumptions about binaries without attempting to equalize or neutralize bias.

1.7 connection to recent work

2 things

More recent prompting methods perpetuate a categorical way of thinking about bias: in this case, that bias is consious or unconscious. For example, [Kaneko et al. 2024, Dong et al. 2024] use text generation methods to explore so-called implicit or unconscious social bias. They craft prompts that require LLMs to explain their reasoning (CoT) or use "indirect probing". The idea is that LLMs, like humans, contain implicit biases, which can be brought out through explicit prompting.

These methods are effective. The models tend to perform with less biases when prompted.

But this approach, like WEAT, collapses bias into a single measure, overlooking the particular effects of that bias (stereotype, denigration, underrepresentation). It emphasizes the model of "conscious/unconscious" as the primary way of looking at bias, rather than the types of bias. The potential effect is to bury bias deeper into the system.

Bias hides. When we are looking for so-called "unconscious" bias, it will be hidden, sure. But what if we look for it on the surface? What if we use categories that name it by the way it affects certain people: by stereotype, denigration, representation?

The other issue with more recent methods of prompting is the problem-

atic way they put the liability of bias detection and migitation onto the user. While these studies frame it as giving the user more control over model performance, it relieves already relatively unregulated AI developers from responsiblity to produce fair and unbiased models, not to mention for producing models that are open [Thakur et al., ACL 2023, Furniturewala et al., EMNLP 2024]. If bias in closed, proprietary models can be detected and evaluated in efficient ways, these developers have less pressure to produce higher quality products.

1.8 conclusion

[should I make the point that male and female vs masculine and feminine]

The assumptions holding up the apparent stability of the binary drive some of the strategies for detecting and mitigating bias—-strategies that approach it as a categorical measure, or attempt to excise or neutralize it. The idea that bias is categorical is descends from slippage between disciplines that misapprehend how social bias is represented in language forms. In this slip, bias, as social stereotype, under-representation, or denigration is discussed in terms of a single measure, of "prior information," which is assumed to be of a particular type. The binary model in which gender is diametrically opposed implies that gender is distinct and stable. It also implies a framework where "equal" is the same as "equitable," as if bias is a zero-sum phenomenon with the goal of attaining

However, a critical look at Queer Studies' theorization of the binary model reveals that what appears to be symmetrical is in fact skewed. This idea is born out by researchers, who find that many of above the attempts used to mitigate or eliminate bias succeed only in terms of hiding that bias [Gonen and Goldberg, Stanovsky et al. 2019]. Rather than a measurement of error or "prior information", social bias ought to be represented a complex and relational phenomenon, which takes into account other variables like the positionality of speakers, context, tone, and other aspects, in language.

As Abeba Birhane's work on "Encoded Values in Machine Learning" [2022] argues, neutrality can and does obscure harmful assumptions that work to "disproportionally benefit and empower the already powerful, while neglecting society's least advantaged." Moving forward requires understanding that equity is not the same as equality, and that what pertains to one group is not equivalent to what pertains to the other. Under those conditions, eliminating bias may have less to do with reduction, and more, perhaps, to do with proliferation.