Some Myths About Bias: A Queer Studies Reading Of Gender Bias In NLP

Anonymous ACL submission

Abstract

This paper critiques common assumptions about gender bias in NLP. It argues that many existing bias detection and mitigation techniques in NLP reveal a kind of "binary thinking" that goes beyond the gender binary into structures of thought that limit their effectiveness. Drawing its critique from the Humanities field of Queer Studies, this paper demonstrates that binary thinking drives two "myths" in gender bias research: first, that bias is categorical (that it can be measured as either present or absent), and second, that it is zero-sum (that the relations between genders are symmetrical). Due to their operationalization of binary thinking, each of these myths flattens bias into a measure that fails to distinguish between the types of bias and its effects in language. The paper concludes by pointing to methods that resist binary thinking, such as those that diversify and amplify gender expressions.

1 Bias Statement

This paper critiques methods used to study and mitigate gender bias in NLP. It adopts a framework from Nemani et al. (2023) that organizes bias into the categories of "Denigration", "Underrepresentation", and "Stereotype", further elaborated in Section 3. It assumes that bias is inherent to language systems, and it demonstrates how some methods that attempt to excise bias from language not only misunderstand its operation but also miss the opportunity to imagine alternative mitigation strategies.

2 Introduction

This paper analyzes methods for evaluating and mitigating gender bias in Large Language Models (LLMs) by drawing from current conceptualizations of gender from the humanities. It argues that mitigating gender bias requires understanding not only the gender binary, but the binary form itself, which has been vigorously theorized in humanities fields that specialize in sex, gender, and sexuality, like Queer Studies. It incorporates domain-specific knowledge from the field of Queer Studies to analyze assumptions about binaries that drive current bias evaluation and mitigation methods.

I have chosen the field of Queer Studies as the foundation for my critique because this field offers a deep analysis of how binary forms determine power structures and delimit what can and cannot be represented within them. This analysis of the binary as an ideological structure goes beyond the contributions typically associated with Queer Studies, which is is Gender Performativity, the notion that gender is a social and behavioral phenomenon (Butler, 1990). Since the development of this theory, which inaugurated the field of Queer Studies in the early 1990s, 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, 2021). At the same time, Queer Studies and related fields, such as Intersectional Feminism, Black Feminist Studies, and Trans Studies have continued to debate and problematize the sex/gender division and how it intersects with other aspects of identity like race and class (Amin, 2022; hooks, 2000; Muñoz, 2009; Klein and D'Ignazio, 2024a).

This paper 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 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, 1993; Sedgwick, 1990; Halberstam, 1998; Amin, 2022).

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. Those that do mention binaries, largely do so in the context of gender binary, i.e., male/female (Hitti et al., 2019; Nemani et al., 2023; Klein and D'Ignazio, 2024b) [1]

To fill that gap, this paper argues that the binary, as a form of thinking that encodes power relations between two terms (and what is excluded from them), implicitly structures the conceptualization of bias in NLP. I expore two "myths" about bias: (1) that bias is categorical, and (2) that bias is zero-sum. These myths reveal that some driving assumptions behind bias evaluation and mitigation: that bias is of a single type, and that equity is the same as equality.

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 prompting, gender swapping and Counterfactual Evaluation methods (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] For example, Lauren Klein and Catherine D'Ignazio in "Data Feminism for AI" call to "rethink binaries".

3 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. Similarly, Blodgett et al. (2020) and 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. (2023)'s useful tripartite scheme for categorizing bias. As I demonstrate below, bias often exceeds a dichotomous 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, such as stereotype and 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 harmful suggests that associations between femininity and motherhood are denigrating, without considering the roles of stereotype and underrepresentation in such associations. These conflations lead to mitigation strategies that are less specific to that particular type of bias, and therefore less effective.

4 Queer Studies on Binaries

While bias detection and mitigation methods in NLP aim for an elimination of bias, the Queer Studies field has problematized the idea that inequality can be eliminated from social systems. In this field, much of the debate centers on how forces of stigmatization and oppression operate within larger systems of power and, from within these unjust dynamics, of finding and devloping alternative means of of survival, liberation, and joy (Love, 2009; Butler, 1993; Muñoz, 2009).

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 their first book, *Gender Trouble: Feminism and the Subversion of Identity* (Butler, 1990), but more robustly theorized in their follow up work, *Bodies That Matter: On the Discursive Limits of Sex* (Butler, 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 their 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 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 into an ordered relation. One effect, is to bring its terms into legibility through contrast and opposition. As Queer Studies scholar Sedgwick (1990) 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, 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, appeared as for the purpose of distinguishing against what was then taken to be a perverse and abberant orientation, what Queer Studies scholar Kadji Amin describes "as a normative ballast against homosexuality" (Amin, 2022) In this case, the term "heterosexual," achieves its definition by circumscribing the content of the other term in the binary. Despite this attempt to stabilize and delimit concepts by illustrating a certain symmetry, the terms of the binary are not symmetrically balanced.

The meaning of each term in the binary is determined by the dynamics between what is represented and what is excluded from that binary, what Butler (1993) calls the binary's "necessary outside." Although excluded from the binary, this element enables its operation. For example, in the "heterosexual/homosexual" binary, not only is "heterosexual" defined in contrast to homosexual, but "homosexual" itself is defined against sexualities that are representable from within that schema, what Butler describes as "a domain of unthinkable, abject, unlivable bodies" (Butler, 1993). In other words, the binary gains its definition precisely by what is excluded from its conceptual system.

The binary's apparent symmetry and totalizing power, therefore, masks an underlying imbalance and partiality. However, despite their constraining nature, binaries, in Sedgwick (1990)'s words, remain "peculiarly densely charged with lasting potentials for powerful manipulation". The dimorphic structure of the binary enables a back-and-forth movement between the two terms, opening the potential of rebound and relay. Halberstam (1998) explains that gender, "multiply relayed through a solidly binary system", creates a mixture of expressions that results in gender non-conformity. By vacillating between two poles, masculine and feminine, additional meanings may accrue that disrupt the binary's original exclusions–a topic I will return to in this paper's Discussion.

In the next section, I explore how these aspects of binary thinking, symmetry and totalizing scope, influence two "myths" in two myths that underpin bias evaluation and mitigation techniques in NLP: (1) that bias is categorical, and (2), that bias is zero-sum.

5 Myth 1: Bias is Categorical

The fist myth is that bias is categorical: that it can be measured by ascribing a score between two values, for instance, between yes/no or present/absent. To demonstrate this effect, I focus on an influential bias evaluation technique, The Word-Embedding Association Test (WEAT) (Caliskan et al., 2017) as well as more recent text generation methods based on prompting. These methods, I argue, collapse and reduce the type of bias (i.e. stereotype, representation, denigration) into a single score. By overlooking the specific type of bias and how it operates against the other types, the downstream effect is that biases remain embedded in language forms.

The myth that bias is categorical begins with a subtle conflation between bias in machine learning and bias in social discrimination, which happens at the outset of WEAT's study. Here, 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 machine learning, bias is a single measure that captures the accuracy and correctness of model output, and it is calculated by subtracting the true value of an output from its expected value. By contrast, in social contexts, bias indicates a pre-conception about a person that is based on aspects of that person's identity and/or physical traits. According to the WEAT authors, bias as a measure of "prior information" is summarily transferred into an indicator of "lead[ing] to harmful behavior" (Caliskan et al., 2017). Crucially, this move assumes that bias is equivalent to one particular type of bias, to denigration, which either ignores other types of bias, like stereotypes and under-representation, or collpases these into a single category.

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 photos of people with one of two labels, such as "fat" or "thin." Then, in a subsequent round of the test, subjects will categorize pleasant or unpleasant words using "good" or "bad." Finally, the test runs for two more rounds with similar prompts in content and structure, except with the response keys switched between the fat/thin and good/bad choices. 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]. 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 IAT's approach toward bias as a categorical value, such as present/absent, effectively imposes an evaluative measure on top of a detection one. The extent to which 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.

One example demonstrates a downstream effect, where bias as underrepresentation becomes conflated with denigration. In a study using word vectors, names that are overrepresented exhibit a higher positivity score, while those that appear fewer times show a negative score (**?**). Here, the frequency 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).

The WEAT metric's development, and particularly the way it adopts concepts from across disciplinary understandings, conceptualizes bias with the effect of limiting the kinds of results bias evaluation techniques can achieve. This is a significant effect for a metric that has influenced the development of other 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 binary thinking that drives vector-based evaluation methods also appears in more recent methods like prompting. These methods use prompt engineering to explore so-called "implicit" or "unconscious" social bias (Kaneko et al., 2024; Dong et al., 2024). By requiring LLMs to explain their reasoning (Chain of Thought or CoT), or through the use of "indirect probing," the idea is that LLMs, like humans, can reveal implicit biases.

While these prompting methods are more successful than vector-based ones, which are proven to be ineffective for measuring downstream bias (Gonen and Goldberg, 2019), they nonetheless impose a binary that constrains their potential. Because these methods impose a binary of conscious/unconscious on the data that they model, they obscure the effect that bias has on identity groups and effectively outsource responsibility for reducing bias. Labelling bias as unconsious overlooks the *explicit* effects of bias, such as underrepresention or denigration, and focuses instead on implicit bias, which is presented as endemic or naturally occuring to the model. As Kaneko et al. (2024) assert, "CoT encourages an LLM to be aware of its hidden biases and articulate a fair thinking process, thus leading to bias mitigation." With this conception of bias as endemic, the responsibility shifts to the user to mitigate the bias, thus relieving model developers, who already encounter low levels of regulations and legal incentives, from social responsiblity. It is worth noting that prompting also reduces the incentive to produce open models, as proprietary models can be evaluated without access to underlying parameters (Thakur et al., 2023; Furniturewala et al., 2024).

6 Myth 2: Bias is Zero-Sum

Rallying all of bias into a categorical label like "present/absent" or "conscious/unconcious" 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. I now move to bias mitigation techniques that build on this premise in the assumption that bias is zero-sum–that it can be manipulated to achieve equality between the sexes.

Another word vector-based technology, "DeBias," is a mitigation strategy that attempts to deduct 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. Depending on whether terms are gender specific or gender neutral ("gal" and "guy" are gender specific, while "programmer" and "babysitter" are gender neutral), those terms are either "equalized" or "neutralized": terms that are neutralized have values closer to zero in the gender subspace, while terms in the equalized set are made equidistant from the gender neutral terms. 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 DeBias shows that a gender subspace cannot be extracted from word vectors like thread from a cloth. Gonen and Goldberg (2019) in particular claim that the results are "superficial," explaining 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". 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).

However, I argue, not all associations have the same effect. Not all stereotypes are harmful in themselves, and sometimes, stereotypes can be descriptive without being delimiting. For example, Gonen and Goldberg (2019) explain, that terms like "math" and "delicate", "have strong stereotypical gender associations" that "reflect on, and are reflected by, neighbouring words". In its association to femininity, the term "delicate" may refer to pleasantnes, subtlty, sensitivity; or, it can refer to weakness or sickness. None of these associations are harmful in themselves. The harm comes from using these latter associations as a basis for further associations that delimit or demean. For example, if the association to weakness marks femininity as needing of protection, or place it within patriarchal notions of control, then the association

is derogatory. Contrast these associations to those that accompany the word pair, "bachelor/spinster". As Devinney et al. (2022) explain, the term "spinster is pejorative while bachelor is not," pointing out that "there is no such thing as a spinster's degree." Close attention to the particular type of bias would help to explain which kinds of associations are harmful and how they should be handled.

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.

These methods have in common the assumption that gender is a zero-sum phenomenon. 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.

7 Discussion

This paper has shown some ways that the binary thinking influences methodological choices for studying bias in NLP. Binaries are totalizing, reducing all complexity into a categorical measure, such as the collapse of different types of bias into a measure of "prior information". They are also symmetrical, placing its terms within a stable opposition so that gendered words can be equalized or neutralized.

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). This charge comes from within the polarizing forces of the binary itself. These polarizing forces are precisely what, Halberstam (1998) explains, enables "gender's very flexibility and seeming fluidity". They can be manipulated to resist their dimorphic constraints of the binary form.

Some recent work in NLP explores this potential through the strategy of bias amplification. This strategy harnesses stereotype to its advantange, to amplify (rather than reduce) stereotype in a model's training dataset. In "Fighting Bias with Bias," Reif and Schwartz (2023), following the work of Stanovsky et al. (2019), include 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).

The strategy of "amplifying bias" harness the binary form without falling into the traps of binary thinking, that is, to equalize or neutralize the terms of the binary. Rather, it opens the possibility to reformulate the binary, a notion well-explored in Queer Studies, particularly in the context of gender non-conforming subjects. Halberstam (1998) offers the example of a masculine-presenting–though not quite female–person:

> What if a biological female who presents as butch, passes as male in some circumstances and reads as butch in others, and considers herself not to be a woman but maintains distance from the category 'man'? For such a subject, identity might be best described as a process with multiple sites for becoming and being. To understand such a process, we would need to do more than map psychic and physical journies between male and female within queer and straight space; we would need, in fact, to think in fractal terms and about gender geometries.

Halberstam's phrase "gender geometries" recalls the "gender subspace" of De-Bias methods. But instead of equalizing or neutralizing the terms of

binary, or in Halberstam's words, to "map… between male and female," he suggests new ways of thinking about those forms and how they are traditionally composed. Perhaps, this means fracturing (or refracting) what has been considered to be wholly or stably "male" or "female" into distinct expressions of behavior, presentation, and self-perception.

8 Conclusion

The binary model implies a framework where everything can be contained within its scope, and where equal is the same as equitable. However, a critical look at Queer Studies' theorization of the binary model reveals that what appears to be stable and symmetrical is in fact skewed. The binary operates through forces of totalization and contrast that places its terms into precarious balance.

Rather than a measurement of error, gender bias ought to take into account the type of bias, such as stereotype, underrepresentation, and denigration, and how these emerge in language. It also might consider the possiblities for working within constraints in order to push their boundaries beyond their traditional forms. In other words, the binary's very constraints–the rigidity of its structure and polarizing forces–can be turned to its potential. Under these conditions, eliminating bias may have less to do with reduction, and more, perhaps, to do with proliferation.

Limitations

The scope of this paper is limited to gender bias and to word vector-based techniques for studying this type of bias. Future work might apply its methods toward studying other types of social bias, such as bias in race and religion. Future work might also lend a deeper attention to bias evaluation and mitigation techniques that are not considered here, or considered briefly, such as prompting, gender swapping, and coreference resolution.

Another limitation is the gender binary itself. This paper focuses on the binary form and does not explicitly consider nonbinary gender identities or research on bias on these identities in NLP. The question of nonbinary representation is a complex one, particularly in how this representation engages a binary schematic. It is the position of this author that the topic of nonbinary representation is urgent and merits dedicated focus in future work.

References

Claire Ainsworth. 2015. Sex redefined. *Nature*, 518(7539):288–291. Publisher: Nature Publishing Group.

Kadji Amin. 2022. We are All Nonbinary: A Brief History of Accidents. *Representations*, 158(1):106–119.

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: from allocative to representational harms in machine learning.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. *arXiv preprint*. ArXiv:2005.14050 [cs].

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Online. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *arXiv preprint*. ArXiv:1607.06520 [cs].

Judith Butler. 1990. *Gender Trouble: Feminism and the Subversion of Identity*. Routledge. Google-Books-ID: gTbbCgAAQBAJ.

Judith Butler. 1993. *Bodies that Matter: On the Discursive Limits of "sex"*. Psychology Press. Google-Books-ID: ZqiIgwQiyFYC.

Aylin Caliskan, Pimparkar Parth Ajay, Tessa Charlesworth, Robert Wolfe, and Mahzarin R. Banaji. 2022. Gender Bias in Word Embeddings: A Comprehensive Analysis of Frequency, Syntax, and Semantics. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 156–170. ArXiv:2206.03390 [cs].

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186. ArXiv:1608.07187 [cs].

Tessa E S Charlesworth, Kshitish Ghate, Aylin Caliskan, and Mahzarin R Banaji. 2024. Extracting intersectional stereotypes from embeddings: Developing and validating the Flexible Intersectional Stereotype Extraction procedure. *PNAS Nexus*, 3(3):pgae089.

Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2022. Theories of "Gender" in NLP Bias Research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 2083–2102, New York, NY, USA. Association for Computing Machinery.

Xiangjue Dong, Yibo Wang, Philip S. Yu, and James Caverlee. 2024. Disclosure and Mitigation of Gender Bias in LLMs. *arXiv preprint*. ArXiv:2402.11190 [cs].

Shaz Furniturewala, Surgan Jandial, Abhinav Java, Pragyan Banerjee, Simra Shahid, Sumit Bhatia, and Kokil Jaidka. 2024. "Thinking" Fair and Slow: On the Efficacy of Structured Prompts for Debiasing Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 213–227, Miami, Florida, USA. Association for Computational Linguistics.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. *arXiv preprint*. ArXiv:1903.03862 [cs].

Anthony G. Greenwald, Debbie E. McGhee, and Jordan L. K. Schwartz. 1998. Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74(6):1464–1480. Place: US Publisher: American Psychological Association.

Jack Halberstam. 1998. *Female Masculinity*. Duke University Press.

Yasmeen Hitti, Eunbee Jang, Ines Moreno, and Carolyne Pelletier. 2019. Proposed Taxonomy for Gender Bias in Text; A Filtering Methodology for the Gender Generalization Subtype. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 8–17, Florence, Italy. Association for Computational Linguistics.

bell hooks. 2000. *Feminist Theory: From Margin to Center*. Pluto Press. Google-Books-ID: uvIQbop4cdsC.

Janet Shibley Hyde, Rebecca S. Bigler, Daphna Joel, Charlotte Chucky Tate, and Sari M. van Anders. 2019. The future of sex and gender in psychology: Five challenges to the gender binary. *American Psychologist*, 74(2):171–193. Place: US Publisher: American Psychological Association.

Daphna Joel. 2021. Beyond the binary: Rethinking sex and the brain. *Neuroscience & Biobehavioral Reviews*, 122:165–175.

Masahiro Kaneko, Danushka Bollegala, Naoaki Okazaki, and Timothy Baldwin. 2024. Evaluating Gender Bias in Large Language Models via Chain-of-Thought Prompting. *arXiv preprint*. ArXiv:2401.15585 [cs].

Lauren Klein and Catherine D'Ignazio. 2024a. Data Feminism for AI. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pages 100–112, New York, NY, USA. Association for Computing Machinery.

Lauren Klein and Catherine D'Ignazio. 2024b. Data Feminism for AI. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 100–112, Rio de Janeiro Brazil. ACM.

Heather Love. 2009. *Feeling Backward: Loss and the Politics of Queer History*. Harvard University Press.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On Measuring Social Biases in Sentence Encoders. In *Proceedings of the 2019 Conference of the North*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2022. An Empirical Survey of the Effectiveness of Debiasing Techniques for Pre-trained Language Models. *arXiv preprint*. ArXiv:2110.08527 [cs].

José Esteban Muñoz. 2009. *Cruising Utopia: The Then and There of Queer Futurity*. NYU Press. Google-Books-ID: f1MTCgAAQBAJ.

Praneeth Nemani, Yericherla Deepak Joel, Palla Vijay, and Farhana Ferdousi Liza. 2023. Gender Bias in Transformer Models: A comprehensive survey. *arXiv preprint*. ArXiv:2306.10530 [cs].

Yuval Reif and Roy Schwartz. 2023. Fighting Bias with Bias: Promoting Model Robustness by Amplifying Dataset Biases. *arXiv preprint*. ArXiv:2305.18917 [cs].

Eve Kosofsky Sedgwick. 1990. *Epistemology of the Closet, Updated with a New Preface*. University of California Press. Google-Books-ID: KMhUa25EPkIC.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating Gender Bias in Machine Translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Himanshu Thakur, Atishay Jain, Praneetha Vaddamanu, Paul Pu Liang, and Louis-Philippe Morency. 2023. Language Models Get a Gender Makeover: Mitigating Gender Bias with Few-Shot Data Interventions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 340–351, Toronto, Canada. Association for Computational Linguistics.

Austin van Loon, Salvatore Giorgi, Robb Willer, and Johannes Eichstaedt. 2022. Negative Associations in Word Embeddings Predict Anti-black Bias across Regions–but Only via Name Frequency. *Proceedings of the ... International AAAI Conference on Weblogs*

and Social Media. International AAAI Conference on Weblogs and Social Media, 16:1419–1424.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. *arXiv preprint*. ArXiv:1804.06876 [cs].