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

**Filipa Calado**
School of Information, Pratt Institute
New York City, USA
fcalado@pratt.edu

## Abstract

This paper critiques common assumptions about gender bias in NLP, focusing primarily on word vector-based methods for detecting and mitigating bias. It argues that these methods assume a kind of "binary thinking" that goes beyond the gender binary toward a conceptual model that structures and limits the effectiveness of these techniques. 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, measuring bias in terms of presence/absence, and second, that it is zero-sum, where the relations between genders are idealized as symmetrical. Due to their use of binary thinking, each of these myths flattens bias into a measure that cannot distinguish between the types of bias and their effects in language. The paper concludes by briefly pointing to methods that resist binary thinking, such as those that diversify and amplify gender expressions.

## 1 Bias Statement

This paper adopts a framework from Nemani et al. (2023) that organizes bias into the categories of "denigration", "underrepresentation", and "stereotype", within the larger category of "representational harms," 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 focus on a binary structure of thought that miss the opportunity to imagine alternative mitigation strategies.

## 2 Introduction

This paper analyzes methods for evaluating and mitigating gender bias in NLP, focusing primarily on word vector-based methods, 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 choose 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. My 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).

This paper considers the binary as not just a way of categorizing and understanding 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 positions 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 focuses on word vector-based meth-



*Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 338–346

ods, as well as some prompting and gender-swapping methods, furthering areas of NLP research that are already robust with critiques of bias detection and mitigation techniques. While many studies have pointed out how such methods are ineffective (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, 2024a).[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 demonstrate this point by introducing two "myths" about bias: (1) that bias is categorical, and (2) that bias is zero-sum. I argue that these myths drive some foundational assumptions behind bias evaluation and mitigation techniques: that bias can be reduced to one kind of effect, which is harm, and that seeking equality between social groups creates social equity.

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 the binary as a conceptual model, and how it necessitates certain exclusions to reinforce its apparent stability. 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). While I briefly mention other methods, such as those that use prompt engineering and gender swapping (Zhao et al., 2018; Meade et al., 2022; Nemani et al., 2023), I focus on word vectors because they offer a close-up view of the semantics that operate within binary structures. Finally, I close by pointing to some promising work in current NLP research that operationalizes the binary model in capacious and productive ways.

[1]One exception to this is Lauren Klein and Catherine D'Ignazio's call to "rethink binaries".

## 3 Gender Bias in NLP

The existing research on gender bias in NLP conceptualizes bias according to certain features and/or effects, such as social stigma, resource allocation, and syntactic structures, among others, which are difficult to map into one totalizing schema. Generally, however, the research defines bias into two kinds: by how it is expressed in language (structural and grammatical expressions), and by its social effects (representational and allocative effects).

Hitti et al. (2019), who examine how bias appears in language, further divide bias into structural and contextual types. Structural bias describes 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 describes bias that results from social and behavioral stereotypes ("Senators need their wives to support them throughout their campaign") (Hitti et al., 2019). Moving from these structural expressions to social effects, Nemani et al. (2023) classify bias by the particular implication that it has for a specific social group, and organizes bias into the categories: "Denigration," "Stereotyping," and "Underrepresentation." Denigration refers to the use of derogatory language such as slurs; stereotyping refers to prejudice about a particular social group; and underrepresentation refers to the relative dearth of information about a particular social group. In a similar schematic, 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 underrepresented or stereotyped.

This paper focuses on bias that has to do with representation, specifically on the semantics of individual words and what they represent about a social group. To describe such effects, it adopts Nemani et al. (2023)'s useful tripartite scheme of "denigration," "stereotype," and "underrepresentation." As demonstrated below, bias often exceeds a dichotomous measure, so that having multiple categories will yield more precise and illustrative analysis.

As such, this work offers a critique of current research on bias which does not distinguish between these categories to the effect of conflating one with another, such as stereotype with denigration. This oversight, which I argue is attributable binary thinking, collapses different types of bias within one

reductive frame. For example, the common assumption that all bias is harmful suggests that associations between femininity and motherhood are denigrating, without considering the descriptive functions and 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 both the fields of NLP and Queer Studies admit that bias cannot be completely eliminated from social systems—that there is no such thing as perfect equality— Queer Studies has gone further in exploring the contradictions that underlie the ideals of social egalitarianism. In this field, much of the debate centers on how forces of stigmatization and oppression operate within larger systems of power, and of finding and developing alternative means of survival and practices of liberation from within these unjust dynamics (Love, 2009; Butler, 1993; Muñoz, 2009). The extent to which Queer Studies has problematized structures of power relating to gender in particular, I argue, offers a useful resource for theorizing gender bias evaluation and mitigation methods in NLP.

One enticing problematic for Queer Studies has been the gender binary and binary structures generally. The field-forming deconstruction of the gender binary 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 (Amin, 2022); heterosexuality, in other words, appeared as for the purpose of distinguishing against homosexuality, in what Queer Studies scholar Amin (2022) describes "as a normative ballast against homosexuality". In this case, the term "heterosexual," achieves its definition by circumscribing the content of the other term in the binary, the "homosexual," which was then considered to be a perverse and aberrant sexuality. Despite this attempt to stabilize and delimit sexuality by suggesting 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." 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 unrepresentable from within that schema, what Butler describes as "a domain of unthinkable, abject, unlivable bodies" (Butler, 1993). For Butler, this "outside" is "necessary" because 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, this dynamic also opens the potential for gender non-conformity. Despite their constraining nature, binaries are, in Sedgwick (1990)'s words, "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", can create a mixture or layering 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 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 as 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 some more recent text generation methods based on prompting. These methods, I argue, display a tendency to collapse and reduce the type of bias (i.e. stereotype, representation, denigration) into a single score. By overlooking the specific category of bias and how it operates against other categories, the downstream effect is that biases remain embedded in language forms.

The myth that bias is categorical begins with a subtle conflation of "bias" between machine learning and social discrimination contexts. I argue that this conflation, which is common and indeed drives some bias evaluation and mitigation research, appropriates the definition of bias from a social discrimination context to a machine learning one. One notable example appears at the outset of the WEAT study, an influential word-embedding method for studying social bias in word associations. 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). By emphasizing bias which "lead[s] to harmful behavior," the WEAT authors prioritize one effect of bias, that is, denigration, over other effects (Caliskan et al., 2017). This move, which summarily transfers bias from a social domain to a computational one, leaves out discussion about other types of bias, such as stereotype and underrepresentation, which have different effects, and how to address those effects.

The WEAT indicates biases as a single measure that represents implicit preference or aversion. Adapted from social psychology's Implicit Association Test (IAT) (Greenwald et al., 1998), the test subject will first 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, 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 indicates an implicit preference and a longer distance indicates an implicit 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 WEAT, subsequently, in another appropriation from a social domain, from social psychology to machine learning, pinpoints intra-group prejudice within vector space. This transaction takes a categorical quality transforms it into a numerical score, indicating the strength of association. While this method may be useful for indicating implicit preference or aversion, the extent to which an association can be detected does not indicate the harmfulness of that association, not to mention its particular quality or effect—having to do with stereotype, representation, or denigration, for example.

The conflation between between bias and harm, which is common in bias mitigation research, associates the presence of something with its effect. WEAT, for example, correlates word associations to implicit preferences and aversions.

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 (Wolfe and Caliskan, 2021). Here, the frequency of certain group names, those of typically minority groups, has a derogatory 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 are still constrained by categorical assumptions. Because these methods impose a binary of conscious/unconscious on the data that they model, they not only obscure the specific type of bias but also effectively outsource the responsibility for reducing bias. Labelling bias as unconscious overlooks the *explicit* effects of bias, such as underrepresentation or denigration, and focuses instead on implicit bias, which is presented as endemic or naturally occurring, so-called "hidden biases" by one group of researchers (Kaneko et al., 2024). The combination of prompting methods along with this conception of bias as endemic shifts the responsibility to the user to mitigate the bias, thus relieving model developers, who already encounter low levels of legal regulations and incentives for building socially responsible models. 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 types of bias into a categorical label like "present/absent" or "conscious/unconscious" 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 that are equalized are made equidistant from the gender neutral terms. The developers explain that, "after equalization babysit would be equidistant to grandmother and grandfather and also equidistant to gal and guy, but presumably closer to the grandparents and further from the gal and guy" (Bolukbasi et al., 2016).

However, 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 be-

tween '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 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, neighboring words". In its association to femininity, the term "delicate" may refer to pleasantness, subtlety, 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 femininity. 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 harmful. Compare that indirect association of harm to more direct associations that accompany the word "spinster," especially when compared to its masculine counterpart, "bachelor." 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 if they ought to be mitigated.

The idea that gendered terms can operate "neutrally" or "equally" across contexts influences other bias mitigation techniques which are based in gender swapping (Zhao et al., 2018). These methods generally take a single dataset and swap out gender terms, such as "actor" for "actress," and assess differences across outputs. 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 may carry connotations that do not make them equivalent or able to be substituted one for the other.

Rather than a zero-sum phenomenon, 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 ways that bias operates across embedding space.

In the next section, I offer a starting place for working within the constraints of the binary structure to mitigate gender bias in language.

## 7 Discussion

This paper has shown some ways that the binary thinking influences methodological choices for studying bias in NLP, particularly those related to word vector technology. 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 which, according to Halberstam (1998), enables "gender's very flexibility and seeming fluidity." In other words, as Queer Studies scholars argue, the dimorphic constraints of teh binary form can be manipulated to resist the binary's very rigidity.

Some recent work in NLP explores this potential through the strategy of bias amplification. This strategy harnesses stereotype to its advantage, 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 occupations ("doctor"), will result in gendering "doctor" as female (or alternatively, describing a male gender as "pretty" (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" harnesses the binary form without falling into the trap 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-identifying—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 journeys between male and female within queer and straight space; we would need, in fact, to think in fractal terms and about gender geometries.

Here, Halberstam's use of geometric and graphical imagery evokes the word vector methods discussed previously. However, rather than conceptualizing a "gender subspace," where the binary aspires toward ideal symmetry of equalizing or neutralizing its terms, Halberstam's "gender geometries" seeks another use of the binary. Perhaps, this means fracturing (or refracting) what has been considered to be wholly and firmly "male" or "female," and exploring new compositions created from them.

## 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 possibilities 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 word vector-based techniques for studying gender bias. I prioritize word vector techniques for two reasons: first, because they enable a close-up view of semantics for studying binary structures; and second, due to the limitations of space. Future work might 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, among others.

Another limitation is the gender binary itself. This paper focuses on the binary form from within a Queer Studies perspective and does not explicitly consider nonbinary gender identities. Future work might incorporate theorizing about nonbinary identity and how it interacts with other aspects of identity, like race and class, which has been vigorously theorized in fields like Trans Studies, Intersectional Feminism, and Black Feminist Studies (Amin, 2022; hooks, 2000; Muñoz, 2009; Klein and D'Ignazio, 2024b).

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 *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 100–112, Rio de Janeiro Brazil. ACM.

Lauren Klein and Catherine D'Ignazio. 2024b. 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.

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.

Robert Wolfe and Aylin Caliskan. 2021. Low Frequency Names Exhibit Bias and Overfitting in Contextualizing Language Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 518–532, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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].