Some Myths About Bias

A Queer Studies Reading of Bias Evaluation and Mitigation Techniques in NLP

Filipa G. Calado, PhD

Assistant Professor, Pratt Institute, School of Information, fcalado@pratt.edu

This paper critically examines gender bias in large language models (LLMs) by integrating concepts from Queer Studies, particularly the theory of gender performativity and the critique of binary models. It argues that many existing bias detection and mitigation techniques in Natural Language Processing (NLP), such as the Word Embedding Association Test (WEAT) and gender swapping methods, rely on outdated conceptualizations of gender, which take for granted binary gender as a symmetrical and stable form. Drawing from Queer Studies, the paper highlights three "myths" about gender bias: that bias can be excised, that it is categorical, and that it can be leveled. The paper critiques common approaches like word vector-based metrics and counterfactual evaluations for their inherent assumptions of binary gender roles, which impairs their potential to recognize and counter gender bias. The paper concludes by suggesting that bias mitigation in NLP should focus on amplifying diverse gender expressions and incorporating non-binary perspectives, rather than attempting to neutralize or equalize them. By reworking that which is outside the binary, against which the binary defines itself, one 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, word embeddings.

ACM Reference Format:

Filipa G Calado. 2024. Some Myths About Bias: A Queer Studies Reading of Bias Evaluation and Mitigation Techniques in NLP. ACM Conference on Fairness, Accountability, and Transparency, Athens, Greece, June 23-26, 2025.

1 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, Queer Studies, and Trans Studies. It incorporates domain-specific knowledge from these fields to analyze how embedded assumptions about gender binaries drive current bias evaluation and mitigation methods.

The field of Queer Studies in particular has done much to enrich and complicate the understanding of gender: what it is, how it operates, according to what structures and imperatives. For example, the theory of Gender Performativity, which inaugurated the field in the early 1990s, influences the common perception today that gender is a socially constructed phenomenon, which is determined and made visible through social norms and behaviors, rather than the sexed body [Butler 1992]. Since then, the distinction between gender as a social operation and sex as a physical embodiment, and the subsequent disillusion 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 as well as the implications of having a politics based on gender identity [Sedgwick 1990, Edelman 2004, Love 2009]. In addition to those developments, theorizing around intersectionality, or the ways in which gender intersects with race, class, ability, sexuality, and other aspects of identity, has risen to prominence and is now accepted as a standard paradigm for critical analysis in the humanities [hooks 2000, Munoz 2009].

Despite the acceptance of concepts like gender performativity, nonbinary gender, and intersectionality in Humanities scholarship, these generally do not influence developments in Computer Science fields like NLP. This paper considers how such concepts might influence the study of gender bias, and particularly its evaluation and mitigation techniques. It reviews popular techniques, particularly those that deploy word vector-based metrics 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 as an essential component of their operation [Zhao et al. 2018, Meade et al. 2022, Nemani et al. 2023]. It elaborates how many of these methods, which have and continue to significantly influence anti-bias development, perpetuate a conceptualization of gender that centers on a binary model that is self-limiting.

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. To fill that gap, this paper identifies three myths about gender bias that turn on a fundamental misunderstanding of how binaries are formulate and how they function: (1) that bias is excisable, (2) that it is categorical, and (3) that it can be leveled.

In its application of Queer Studies to NLP, this paper attempts to answer calls for more interdisciplinary work in NLP research [Klein and D'Ignazio 2024, Birhane et al. 2022, Devinney et al. 2022]. 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 structures, laying out the ways that binary models of organization necessitate certain exclusions and disavowals that eventually 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 and gender swapping strategies. 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.

2 LITERATURE REVIEWS

2.1 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]. By contrast, Nemani et al. [2023] classify bias by the particular kind of effect it has on social groups, organizing them 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.

2.2 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.[1]

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, subjects do not have a gender, per se, but they express one by behaving according to certain social expectations.

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 hierarchical social order [Butler 1993].

As many Queer Studies scholars point out, social hierarchies are reinforced through the imposition of binaries, such as "male/female" and "heterosexual/homosexual. " The apparent symmetry of the binaries, where two terms seem to have equal semantic weight, mask an instability. In another seminal book in Queer Studies, *The Epistemology of the Closet*, Eve Kosofsky Sedgwick [1990] explains that in the binary "heterosexual/homosexual", the term "heterosexual" is not simply symmetrical or subordinated to "homosexual," but rather, depends "homosexual" for its meaning through "simultaneous subsumption and exclusion." In other words, one term is not simply juxtaposed to another, but rather relies on the other for its definition which is achieved via exclusion and circumscription. Furthermore, in order for a binary to operate meaningfully, something must be excluded from the model, an unrepresentable element that Butler calls the binary's "necessary outside." For example, the binary "heterosexual/homosexual," produces "heterosexual" against the "homosexual, " which itself is produced against "a domain of unthinkable, abject, unlivable bodies" [1993]. In this conceptualization, the term "heterosexual," gains its definition precisely by what is excluded from that conceptual system, the "homosexual," which itself is defined against a sexuality that is not representable from within that schema.

In Queer Studies, then, binaries are theorized as constraining structures that circumscribe certain roles into legibility through the mechanism of exclusion. However, despite their constraining nature, binaries, in Sedgwick's words, remain "*peculiarly* densely charged with lasting potentials for powerful manipulation" – a topic I will return to in this paper's conclusion [1990].

---

1. 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].

3 SOME MYTHS ABOUT BIAS

3.1 Myth 1: Bias is Excisable

One approach for bias mitigation aims to reduce bias from LLM training datasets. Due to the indiscriminate nature of large-scale data gathering methods like web crawling, data filtering is always necessary to some degree. However, when filtering for biased language, it is important to consider the ways that harms and denigration engage with issues of minority group representation. Not doing so runs the risk of "disproportionately remov[ing] text from and about minority individuals," as Dodge et al. [2021] point out.

Accounts of removing bias via filtering show that such strategies do not take the nuances around language context into account. For example, the "c4" dataset [AllenAI 2021], a collection of Common Crawl data dumps that are used to train transformer models like T5, the GPT family, and LaMDA [Thoppilan et al. 2022, Bender et al. 2021, Raffel et al. 2023], infamously uses the "List of Dirty, Naughty, Obscene or Otherwise Bad Words" to filter out discriminatory and sexualized content [LDNOOBW 2012]. The list, which is also available as a JavaScript software package called "naughty-words," focuses primarily on terms associated with online porn, like "bondage" and "camgirl," with others referring to sexual and racial identities, like "bulldyke" and "darkie," and those that describe body parts, like "butt."

While some terms, like "butt," are neutral descriptors that are not in themselves discriminatory or sexualized, many of these terms can carry highly offensive meanings depending on who speaks them, to whom, and for what purpose. The term "bulldyke," for example, although a pejorative term for a masculine-presenting lesbian woman, has been reclaimed by some lesbians that identify with masculine gender expression.[2] Web pages that contain "bulldyke" may therefore describe the meaning of this term from outside of a mainstream and discriminatory point of view. Automating the removal of this content thus runs the risk of excluding terms that, as Bender et al. [2021] explain, "reclaim slurs and otherwise describes marginalized identities in a positive light."

3.2 Myth 2: Bias is Categorical

Besides word filtering, attempts to mitigate bias have leveraged metrics based on word vectors, such as WEAT (The Word-Embedding Association Test [Caliskan et al. 2017]). However, as I demonstrate below, the development of this metric, and in particular the way it engages with related concepts from Social Psychology and Social Sciences, collapses bias into a categorical phenomenon, thus limiting the kinds of results evaluation and mitigation techniques can achieve.

The WEAT Score, developed by Caliskan et al. [2017], combines principles from Social Psychology and Computational Linguistics to measure gender bias in LLMs. From 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." In the IAT, the subject will categorize photos of people with a certain label, such as "fat" or "thin," using their right or left hands which contain a response key [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. Then, for the next two rounds of the test, the response key will switch from one hand to another, and subjects will again categorize words and photos. 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

2. Interestingly, there is debate whether the term originally meant "false man" (*bull* as in false, and *dyke* as in "dick") or "masculine woman" (*bull* as in masculine, and *dyke* as a ridge-like protrusion). See Krantz [1995].

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].[3]

The WEAT takes this idea of social group evaluation into to vector space, using co-sine similarity as a correlative to response time. In doing so, WEAT inherits the binary measurement from its progenitor, and with it, an association of bias as a categorical value, as an evaluative term that is either "good" or "bad." The use of evaluative labels such as "good" or "bad" to detect bias implies that bias is a categorical value, which can be either helpful or harmful, obscuring the particular quality, source, or effect of that bias. Therefore, the AIT's approach toward bias as something that can be represented as good or bad effectively imposes an evaluative measure on top of a detection one. This subtle imposition, as a result, perpetuates a framework for bias detection that fundamentally misses the ways that bias is conceptualized and operationalized in language and practice. That some bias can be judged as harmful in no way indicates the particular harm, in the above case, of fatphobia, and how it might be countered or mitigated.

The slip from bias detection to evaluation is a result of translating a metric from Social Psychology directly into Machine Learning. Another, related slippage occurs in the opposite direction, from Machine Learning into Social Science, with the term "bias" itself. The WEAT authors explain 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, the concept of bias centers on accuracy: it is a measure that affects model performance with regard to the correctness of model outputs. This understanding of bias is then summarily transferred into a Social Science context, where bias indicates a complicated and layered social and relational phenomenon. Rather than a measurement of error, social bias ought to be represented in terms of context, quality, and valence.

Critiques of WEAT reveal downstream effects of this logic, and the difficulty of handling its unexpected results. For example, in another study using word vectors to detect bias, a correlation arises between name frequency and positive or negative associations [Wolfe and Caliskan 2021]. Names that appear often in the training corpus exhibit a higher positivity score, while those that appear fewer times attain a negative score. The effect is to attach a negative association to relatively underrepresented names, such as those from minority groups, thus perpetuating their marginalization. To correct for this result, another study [van Loon et al. 2022] controls for the variable of term frequency. However, the results form that study reveal that this particular "unintuitive aspect of word embeddings…. indicates 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].

3.3 Myth 3: Bias can be Leveled

A related misconception about bias is that it can be leveled, so that gendered terms operate equally across all contexts. Evaluation and mitigation techniques reveal this misconception most starkly in the use of gender swapping, such as in Counterfactual Evaluation and Hard DeBias, among others [Nemani et al. 2023, Bolukbasi et al. 2016].

Counterfactual Evaluation methods measure gender bias by swapping gender terms (from "he" to "she", or "she" to "he", for example) in and assessing their effect on model performance. A similar method uses Winograd-schema style

3. 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].

templates, like Winobias, which use coreference resolution to evaluate a model's association of a particular pronoun with a stereotypical attribute [Zhao et al. 2018].

Because the results of these tests reflect only a change in gender, it seems reasonable to claim that they may be used to measure gender bias. However, these methods do not take into account how gendered terms carry particular connotations that do not make them equivalent, able to 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." This inequivalence is similar to what Blodgett et al. [2021] call the "incommensurable" aspects of certain identity groups, in which two terms are not mutually exclusive. The example they give is between an "American" and a "Latino," which "requires the assumption that a Latino is not an American," when in reality there are millions of Latinos who are also Americans [Blodgett et al., 2021]. Not only do seemingly opposite terms have different connotations, their meanings may not be reducible to an equivalent, or 1:1, relationship.

The supposed equivalent quality of gendered terms is translated into semantic weight in another mitigation strategy, "DeBias," which harnesses word embedding technology to eliminate gendered terms from a model's vector space. Developed by Bolukbasi et al. [2016], this strategy first constructs binaries, what they call "equality sets," of gendered terms, like "grandmother/grandfather," and "gal/guy." Then, it calculates a "gender subspace" or "gender direction" for these equality sets and for gender neutral terms, like "babysitter" and "programmer." Finally, terms which are gender neutral are "Neutralized" by ensuring their values are zero in the gender subspace, while terms in the equality set are "Equalized," or made equidistant from the gender neutral terms. For instance, "if {grandmother, grandfather} and {guy, gal} were two equality sets, then 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].

This method has received some criticism, with Gonen and Goldberg [2019] in particular claiming that the results are "superficial." They explain 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]. The extractive method does not work because word meanings are complexly embedded in such a way that they cannot be isolated and pulled out like a single thread from a cloth. For example, words that carry a specific gender connotation, like "beard," can have unexpected associations even in vector space. While the term "beard," Devinney et al. [2022] explains, generally refers to men, it can also, and ironically, "specifically refer to a woman whom a gay man is dating to hide his sexuality – making it a feminine noun in these cases."

Despite these criticisms, the underlying strategy of using word embeddings continues to influence a distinct trajectory of development for measuring and mitigating bias. For example, both SEAT (The Sentence Embedding Association Test) [May et al. 2019] and SentenceDebias [Liang et al. 2020], expand the use of single-word vector representations to sentence-level representations. As such, they extend the assumption that biased language can be leveled or made equal among groups. By contrast, as I explain in the next and final section, debiasing may benefit from another approach.

4 CONCLUSION

A critical look at Queer Studies' theorization of the binary model reveals that what appears to be distinct, symmetrical, and stable is in fact slippery, skewed, and manifold. I have shown how the assumptions holding up the apparent stability of the binary drive some of the strategies for detecting and mitigating bias—strategies that attempt excise it, approach it as categorical, or level it.

The binary model implies a framework in which "equal" is the same as "equitable," as if bias is a zero-sum phenomenon with the goal of attaining neutrality. However, as Abeba Birhane's work on "Encoded Values in Machine Learning" [2022] aruges, neutrality can and does obscure harmful assumptions that work to "disproportionally benefit and empower the already powerful, while neglecting society's least advantaged."

But this work does not recommend that we leave the binary behind. There are other promising possibilities for handling binaries, also theorized in Queer Studies. Butler, for example, offers a method for reworking a binary's delimiting power. She explains that the "unthinkable outside," which exists to define and circumscribe the binary, can be fashioned into a powerful resource. She gives the example of the term "queer," which previously was a term of denigration that has since been reclaimed, "resignifying the abjection of homosexuality into defiance and legitimacy" [Butler 1993].

There are methods in NLP that take the "unthinkable" and "necessary" outside of established binaries as raw materials for creating datasets to mitigate gender bias. In "Fighting Bias with Bias," Reif and Schwartz [2023] demonstrate a promising approach: amplifying rather than reducing bias in a model's training dataset. They point out that bias reduction techniques are not very effective, that "filtering can obscure the true capabilities of models to overcome biases, which might never be removed in full from the dataset" [Reif and Schwartz 2023]. Instead of reducing, they follow the work of Stanovsky et al. [2019], who intensify bias by including phrases like "the pretty doctor" so that a model will interpret "doctor" to be female. Other approaches take what Devinney et al. [2022] call "trans-inclusive methodologies." 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.

Moving forward requires understanding that equity is not the same as equality. 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.

REFERENCES

Claire Ainsworth. 2015. Sex redefined. *Nature* 518, 7539 (February 2015), 288–291. https://doi.org/10.1038/518288a

AllenAI. 2021. C4.

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

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, March 03, 2021. ACM, Virtual Event Canada, 610–623. https://doi.org/10.1145/3442188.3445922

Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The Values Encoded in Machine Learning Research. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, June 21, 2022. ACM, Seoul Republic of Korea, 173–184. https://doi.org/10.1145/3531146.3533083

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)*, 2021. Association for Computational Linguistics, Online, 1004–1015. https://doi.org/10.18653/v1/2021.acl-long.81

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. https://doi.org/10.48550/arXiv.1607.06520

Judith Butler. 1990. *Gender Trouble: Feminism and the Subversion of Identity*. Routledge.

Judith Butler. 1993. *Bodies that Matter: On the Discursive Limits of "sex."* Psychology Press.

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*, July 26, 2022. 156–170. https://doi.org/10.1145/3514094.3534162

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (April 2017), 183–186. https://doi.org/10.1126/science.aal4230

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*), June 20, 2022. Association for Computing Machinery, New York, NY, USA, 2083–2102. https://doi.org/10.1145/3531146.3534627

Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. 2020. Multi-Dimensional Gender Bias Classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, November 2020. Association for Computational Linguistics, Online, 314–331. https://doi.org/10.18653/v1/2020.emnlp-main.23

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, November 2021. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 1286–1305. https://doi.org/10.18653/v1/2021.emnlp-main.98

Lee Edelman. 2004. *No Future: Queer Theory and the Death Drive*. Duke University Press.

Virginia K. Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. 2024. WinoQueer: A Community-in-the-Loop Benchmark for Anti-LGBTQ+ Bias in Large Language Models. https://doi.org/10.48550/arXiv.2306.15087

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. https://doi.org/10.48550/arXiv.1903.03862

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 (1998), 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

Tony Greenwald, Mahzarin Banaji, Brian Nosek, Bethany Teachman, and Matt Nock. 2011. About the IAT. Retrieved January 22, 2025 from https://implicit.harvard.edu/implicit/iatdetails.html

Saga Hansson, Konstantinos Mavromatakis, Yvonne Adesam, Gerlof Bouma, and Dana Dannélls. 2021. The Swedish Winogender Dataset. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, May 31, 2021. Linköping University Electronic Press, Sweden, Reykjavik, Iceland (Online), 452–459. Retrieved January 22, 2025 from https://aclanthology.org/2021.nodalida-main.52/

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*, August 2019. Association for Computational Linguistics, Florence, Italy, 8–17. https://doi.org/10.18653/v1/W19-3802

bell hooks. 2000. *Feminist Theory: From Margin to Center*. Pluto Press.

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 (2019), 171–193. https://doi.org/10.1037/amp0000307

Daphna Joel. 2021. Beyond the binary: Rethinking sex and the brain. *Neuroscience & Biobehavioral Reviews* 122, (March 2021), 165–175. https://doi.org/10.1016/j.neubiorev.2020.11.018

Andrew Karpinski and James L. Hilton. 2001. Attitudes and the Implicit Association Test. *Journal of Personality and Social Psychology* 81, 5 (2001), 774–788. https://doi.org/10.1037/0022-3514.81.5.774

Lauren Klein and Catherine D'Ignazio. 2024. Data Feminism for AI. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, June 03, 2024. ACM, Rio de Janeiro Brazil, 100–112. https://doi.org/10.1145/3630106.3658543

Susan E. Krantz. 1995. Reconsidering the Etymology of Bulldike. *American Speech* 70, 2 (Summer 1995), 217–221.

LDNOOBW. 2012. LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words. Retrieved January 22, 2025 from https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards Debiasing Sentence Representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, July 2020. Association for Computational Linguistics, Online, 5502–5515. https://doi.org/10.18653/v1/2020.acl-main.488

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. *Proc Int AAAI Conf Weblogs Soc Media* 16, (May 2022), 1419–1424. https://doi.org/10.1609/icwsm.v16i1.19399

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*, 2019. Association for Computational Linguistics, Minneapolis, Minnesota, 622–628. https://doi.org/10.18653/v1/N19-1063

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2022. An Empirical Survey of the Effectiveness of Debiasing Techniques for Pre-trained Language Models. https://doi.org/10.48550/arXiv.2110.08527

José Esteban Muñoz. 2009. *Cruising Utopia: The Then and There of Queer Futurity*. NYU Press.

Praneeth Nemani, Yericherla Deepak Joel, Palla Vijay, and Farhana Ferdousi Liza. 2023. Gender Bias in Transformer Models: A comprehensive survey. https://doi.org/10.48550/arXiv.2306.10530

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. https://doi.org/10.48550/arXiv.1910.10683

Yuval Reif and Roy Schwartz. 2023. Fighting Bias with Bias: Promoting Model Robustness by Amplifying Dataset Biases.

https://doi.org/10.48550/arXiv.2305.18917

Ulrich Schimmack. The Implicit Association Test: A Method in Search of a Construct - Ulrich Schimmack, 2021. Retrieved January 22, 2025 from https://journals.sagepub.com/doi/10.1177/1745691619863798

Eve Kosofsky Sedgwick. 1990. *Epistemology of the Closet, Updated with a New Preface*. University of California Press.

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*, July 2019. Association for Computational Linguistics, Florence, Italy, 1679–1684. https://doi.org/10.18653/v1/P19-1164

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. LaMDA: Language Models for Dialog Applications. https://doi.org/10.48550/arXiv.2201.08239

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*, November 2021. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 518–532. https://doi.org/10.18653/v1/2021.emnlp-main.41

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. https://doi.org/10.48550/arXiv.1804.06876