

Queer Bias in Natural Language Processing: Towards More Expansive Frameworks of Gender and Sexuality in NLP Bias Research

Azure Zhou

Stanford University

1. Introduction

Generative models like OpenAI’s ChatGPT have created much excitement with the potential benefits of their natural language processing (NLP) capabilities for education, industry, and everyday life (*Introducing ChatGPT*, 2022). Alongside such optimism, however, concerns have risen over these models’ disadvantages, from misinformation to social harms. Firstly, large language models (LLMs) derive from massive amounts of unconsented and opaque data collection (Hamilton, 2023; Heikkilä, 2023). Secondly, the industry engages in abusive labor practices (Perrigo, 2023; Hao & Seetharaman, 2023), environmental harms, and poor data documentation that enables biased data to proliferate in downstream tasks (Bender et al., (2021). While much has been said about the biases of high-risk algorithms that govern jurisprudence and finance (Angwin et al., 2022; Mehrabi et al., 2021), the danger also remains ever present in everyday general usage for text generation, web search, and content moderation (Hovy & Spruit, 2016; Pinchevski et al., 2023). Given this potential to “amplify existing biases at unprecedented scale and speed” (Caliskan, 2021, p.1), the machine learning community has focused on measuring and mitigating bias on several fronts: e.g. stereotypes in word embeddings (Bolukbasi et al., 2016; Caliskan et al., 2017), coreference resolution (Zhao et al., 2018), machine translation (Prates et al., 2020), toxicity detection (Dixon et al., 2018; Park et al., 2018) as well as increasing attention to training practices. (Chen et al., 2023). Nonetheless, bias mitigation research has largely centered on race and (binary) gender, while queer identities remain understudied. Today, with widespread public discussion about the best practices of regulating generative models, there is also an opportunity to protect vulnerable groups and create language technologies which affirm and uplift queer individuals.

Owing to the work of Queer in AI, a global, grassroots organization that aims to “raise awareness of queer issues in AI/ML” which has demonstrated the potential of intersectional and community-led participatory design (Queer in AI, 2023), there has been greater opportunity to address the research gaps pertaining to queer bias in NLP systems. Indeed, queer research is increasingly visible at major professional conferences such as the AAI/ACM conference on Artificial Intelligence, Ethics and Society (AIES), where Dennler et al., (2023) organized a participatory workshop to advocate for queer community ownership and participatory processes that shape evaluation of benchmarks, datasets, and model documentation (Dennler et al., 2023). Yet, much work remains to be done. This paper investigates the unique social and computational factors which disincentive research in queer bias and reviews current methods and benchmarks for evaluating such bias. To this end, this paper is laid out as follows: In Section 2, I describe sources of bias and the risks that biased NLP systems pose for queer people. In Section 3, I

synthesize different factors which disincentivize or act as barriers to research in queer bias. In Section 4, I review common frameworks of gender and sexuality in NLP bias research and argue for the need to break out of the academic feedback loop in the theorization of gender and sexuality. Lastly in Section 5, I consider case studies of recent attempts to evaluate queer bias and to debias language models and provide recommendations on how to expand on their frameworks.

2. Harms for queer communities

In “Critically Queer,” Judith Butler (1997) argues that the term queer “has to remain that which is, in the present, never fully owned, but always and only redeployed, twisted, queered from a prior usage and in the direction of urgent and expanding political purposes” (p.19). Employing Butler’s definition, the term “queer” in this paper will refer to people “practicing non-normative gender and sexual identities,” as an umbrella term for non-cisgender or non-heterosexual people, including, but not limited to, lesbian, gay, bisexual, transgender, non-binary, and asexual identities. I will consider gender as it pertains to gender identity rather than sex or gender expression. Gender and sexual identity may be fluid, intersecting, or shift over time, and it is difficult to identify discriminatory language without context. There are many challenges to translating queer lived experience into operational metrics for detecting bias. Here, I use bias to mean difference in model distribution from the true or ideal distribution, which has potential for harm and discrimination based on (protected and unprotected) demographic attributes (Shah et al., 2020). However, creating methodologies to address queer bias is a necessary undertaking due to the risks of NLP systems for queer individuals. I argue for this and review some harm cases below. Following Crawford’s (2017) taxonomy of harms, I distinguish between allocational harms, when systems allocate resources unfairly or perform unequally for different individuals, and representational harms, when systems represent individuals less favorably and inaccurately.

There are many risks associated with AI systems designed without consideration to queer individuals, especially as sociocultural data reproduces the historical and systemic oppression of queer peoples (Tomasev et al., 2021). In just one widely-reported example of how bias in language models surface, a script generated by GPT-3 for a 24-hour Twitch stream related being transgender to a mental illness and suggested that transgender people are “ruining the fabric of society” (Wu, 2023). Queer individuals may additionally face harms which do not always have analogies to race or binary gender, including misgendering and erasure by NLP systems (Dev et al., 2021).

Bias is captured and amplified in many stages of the standard NLP pipeline, from the corpora in which pre-trained models are trained on to the selection of training data to the method of fitting the model itself (Shah et al., 2020). Modern NLP systems use increasingly large text corpora, such as massive web-crawled datasets WebText (Radford et al., 2019) and Colossal Clean Crawl Corpus (C4) (Raffel et al., 2020). These large-scale sociocultural datasets reflect the views and narratives of the people represented, capturing inherent structural inequities, as well as inequities in digital records (Jo & Gebru, 2020). Thus, massive standard texts, widely used in corporate generative models only appear to capture the nuances of the world, but in fact occlude many voices, including those of queer people (Raji et al, 2021). Dev et al. (2021) found that English Wikipedia text contains 15 million mentions of the word he, 4.8 million of she, 4.9 million of they—largely used as the plural pronoun—7.4 thousand of ze, and only 4.5 thousand of

xe—largely referring to the organization Xe (p. 1973). In addition to this inherent distributional skew, web-crawled datasets like C4 have been found to disproportionately remove text associated with minority individuals (Dodge et al., 2021). In their documentation of C4, Dodge et al. (2021) found that mentions of sexual orientations are most likely to be filtered out, in comparison to racial and ethnic identities, and several documents on same-sex relationships were removed (p. 1292).

The direct consequence of this exclusion? Queer individuals are under- and misrepresented in data. Amplification of biases from smaller disparities present in the datasets is a known problem in NLP systems (Zhao et al., 2018). Disproportionate removal of queer texts allows the narratives of queer people in datasets to be largely influenced by non-queer populations, which may lead to representational harm. Additionally, models trained on this data will perform even more poorly when applied to text from and about queer people, an allocative harm. This deterioration in performance is not insignificant. In a recent analysis of sentence completion by LLMs, generated sentences for subjects who are queer was an identity attack 13% of the time, a number which increases for more specific identities, e.g. demisexual (Nozza et al., 2022).

Language models are accordingly prone to the reproduction of heteronormativity, a system which is stabilized by gender binarism and perceives heterosexuality as the naturalized norm (Motschenbacher, 2021). As NLP systems are deployed at scale, this representational harm has the potential to exacerbate sexual and gender disparity. Weidinger et al. (2021) point out the effect LMs may have in downstream applications to “contributing to increasingly homogenous discourse or crowding-out of minority perspectives” (p. 14), echoing concerns by Caliskan (2021) of a feedback loop in which biased outcomes of NLP systems influence societies, which makes bias more prominent in future datasets. Returning to Butler’s (1997) ideas of gender and sexuality, language both reinforces societal roles and is important to the self-construction and performance of gender and sexual identity (Cameron, 2005). For example, the use of languages that emphasize the gender binary is strongly associated with the gender gap in educational attainment (Davis & Reynolds, 2018). The employment of heteronormative language erases asexual and bisexual identities and denies transgender identity, which can severely impact queer people in court decisions (King, 2016). Heteronormative frameworks in media also reinforce stereotypes on queer masculinities (Hindes & Fileborn, 2021).

There are also direct consequences of deploying NLP systems containing queer bias. In a survey of non-binary people, Dev et al. (2021) found that *misgendering* was a concern across several common NLP tasks. Applications can also perpetuate non-binary *erasure* through the failure of text generation to include non-binary people and of speech-to-text or automatic captioning systems to handle neopronouns (Dev et al., 2021, p. 1972). There are also *privacy* concerns that arise when language models leak sensitive information (Weidinger et al., 2021) or infer it across online contexts, which compromise a queer individual’s privacy and safety. Olivia et al. (2020) found that automated content moderation systems consider drag queens to have higher levels of toxicity than white nationalists; however, these technologies may also fail to identify hate speech towards non-binary people. Therefore, content moderation systems online may at the same time *suppress* queer voices and *fail to protect* queer individuals. Lastly, *worse performance* of language technologies when used in hiring, finance, or legal decisions may further the economic divide between non-LBGTQ+ people and LGBTQ+ people, who have faced higher rates of poverty, unemployment, and public benefits use (Medina et al., 2022).

Although NLP systems can provide benefits across many domains, it is necessary to understand the ways in which they can indirectly and directly harm queer communities, reinforcing historical discrimination and perpetuating systemic violence. Through reproducing heteronormativity, these systems also have negative impacts on non-queer individuals and the broader society. These harms can be better understood when making space for the voices of queer individuals in NLP research.

3. Challenges to queer bias research

3.1 Defining bias for queer identities

There are several reasons for which the problem of queer bias in NLP systems has been largely overlooked. Standard approaches in NLP bias define metrics of fairness by comparing between binary attributes—male vs. female, homosexual vs. heterosexual—usually considering only one axis of identity at a time, which follows the paradigm of traditional algorithmic fairness research (Friedler et al., 2019). It is false that an individual must either be homosexual or heterosexual, and to consider queer people as one homogenous group is to ignore the considerable diversity in experiences and challenges faced by different identities within the queer umbrella. As Tomasev et al. (2021) comment, “AI systems may pose divergent risks to different queer subcommunities” (p. 260).

There is a wide range of gender and sexual identities and accompanying labels, which may be fluid, overlapping, and difficult to define. When considering allocational harms, it is therefore necessary to take a granular and intersectional approach to defining bias for queer individuals. Borrowing from the field of algorithmic fairness, multi-group fairness notions (Kang et al., 2021) which consider multiple attributes at once may be equipped to measure and mitigate queer bias for allocative systems. However, multi-group fairness is criticized by Wang et al. (2022) for failure to critically engage with “the substantive differences that intersectionality brings” (p. 337). Wang et al. (2022) further identifies three challenges in moving towards intersectionality: selecting which identities to use, handling the smaller numbers of individuals that lie at intersections, and evaluating as the number of groups increase. In order to make use of their recommendations, however, it is necessary to have robust datasets with labels for identities. This data need is mirrored in bias metric definitions for representational harms, which typically require careful dataset construction: of words, occupations, and actions stereotypically associated with certain demographic attributes; labeled examples of language; or parallel sentence constructions for several groups. Still, data collection for and about queer populations involves several concerns.

3.2 Data collection

When collecting data related to gender and sexual orientation, it is necessary to consider the sensitivity of minority status information and additional burdens involved for queer individuals. This is a concern highlighted by the Office of the Chief Statistician of the United States, which notes significant privacy risks involved with leaking of sexual and gender minority status even after removal of direct identifiers from data (The Office of the Chief Statistician of the United States, 2023). Weidinger et al. (2021) recommend that “sustained mitigation of such harms requires engaging affected groups on fair terms that foreground their needs and interests”

(p.13). Labels must also be self-identified, rather than inferred, and self-identification must be consensually provided, allowing individuals to opt out of the data collection process (Tomasev et al., 2021). That said, those who chose to opt in may not be representative of queer populations more generally.

It is furthermore difficult to identify language which may denigrate, stereotype, or in other ways harm queer individuals in isolation from other social factors, including the identity of the speaker, the social context, and the community's norms (Hovy & Yang, 2021). Instead, the affected groups are best positioned to identify language that may be harmful based on lived experiences. For models which were trained on datasets generated using crowdsourcing, it has been shown that annotator bias can have a significant impact on the generalizability of a model (Geva et al., 2019). Thus, how datasets are constructed and the impacts of annotator positionality on the data collection process deserves more attention.

Advancements in queer bias may also be blocked simply because of the shortage of data authored by and related to queer communities. Even for researchers who incorporate more expansive identities into methodologies, underrepresentation within text corpora still limits results. Correcting for data imbalance by generating synthetic examples may inaccurately conceptualize demographic categories if normative concerns are not considered (Wang et al., 2022). Where mainstream corpora fail to represent queer persons, it may be necessary to introduce additional data collected from and by LGBTQ+ population, as done by Devinney et al. (2020). For massive text corpora which large language models are trained on, reducing bias may require advancement in semi-automated curation methods, which are able to accurately filter out harmful conceptions of queer identities without filtering out queer voices. This again requires careful curation of datasets which can properly capture ideas of heteronormativity and queer exclusion in different contexts.

3.3 Social constructs are not uniform

Another challenge to approaching queer bias is that social constructs and discourses *change over time*, as does the accompanying use of language (Burr & Dick, 2017). This is clear to see in the reclamation of the word “queer” in the past few decades from a pejorative slur to an identity considered empowering (Worthen, 2023). Most methodologies for gender in NLP bias use “pronouns, first names, word pairs, lists of words, or some combination of these strategies to identify referent gender” (Devinney et al., 2022). As the language around queer identity shifts, attempts to categorize words in this manner flattens understanding of queerness, which in addition to being a reductionist approach, may not be applicable to datasets collected from a wider range in time. Therefore, bias definitions must continue to interface with current discourse on sexuality in order to accurately capture how current systems interact with queer individuals and language. In particular, it is necessary to involve queer stakeholders in the NLP development process in order to create NLP systems. As Dennler et al. (2023) argue, transferring ownership of the auditing processes to the impacted communities can help prepare models for better downstream outcomes.

Even so, we must always ask *whose* queer perspectives get included in the auditing process, as each contribution remains *culture-specific*. This paper has centered the English language and Western ideas of gender and sexuality, which do not translate across different or non-Western contexts and is exclusive of other experiences. Stereotypes and stigmas associated with queer people may be more pronounced or arise in different forms due to the historical,

cultural, or political attitudes in different regions. Prerequisite to this, the language used to describe queer experiences in the United States may not even apply to other regions. For example, “kathoei,” an identity used by people in Thailand, does not have a direct translation into Western frameworks, used variably in English to refer to a third sex, effeminate gay men, or transgender women (Käng, 2012).

In countries where identities cannot be expressed without legal repercussions, texts from queer voices and about queer experiences will differ largely from those in countries which guarantee legal protection. It follows that different bias metrics or benchmark datasets may be required to understand how NLP systems treat queer texts across languages and contexts. In terms of access to queer textual data and sensitivity of identity data, as well as the stronger performance of current NLP tools on English vs. low-resource languages (Magueresse, 2020), it is much easier to study concepts of queer bias and heteronormativity in the U.S. However, it is crucial that NLP bias research does not overlook cultures in which language models may learn the most substantial bias and the most stereotyped representations of queer individuals.

4. The standard of gender and sexuality in NLP bias research

In addition to the challenges outlined above, a large barrier to NLP bias research lies in the social frameworks that current researchers intentionally or unintentionally employ. I argue that this is the most difficult hurdle to overcome: that cis- and heteronormative assumptions as the standard in NLP frontiers continues to perpetuate cis- and heteronormative research.

Returning to racial and (binary) gender bias, there is a lack of organized approaches or cohesive metrics and terminology for bias evaluation (Sun et al., 2019; Blodgett et al., 2020; Savoldi et al., 2021). This is partially due to the focus of research on reducing symptoms of bias rather than origins, which can leave the fundamental problem unchanged. For example, even after “de-biasing” word embeddings, bias remains reflected in the distances between words with implicit gender connotations (Gonen & Goldberg, 2019). It is unlikely that one definition of bias will be appropriate for the many different contexts in which language technologies introduce harm; however, it is essential to create a more formal framework on how bias research should be conducted, where assumptions and idealizations are explicitly detailed.

Blodgett et al. (2020) finds that the majority of NLP bias research “fails to engage critically with what constitutes ‘bias’” and is “rife with unstated assumptions about what kinds of system behaviors are harmful, in what ways, to whom, and why” (p. 5454). Noting that the quantitative techniques used to measure or mitigate bias in NLP are often not well-grounded in literature outside of NLP, they recommend NLP bias research to engage more with the literature on the relationships between language and social hierarchies. Savoldi et al. (2021) echoes this criticism, emphasizing that it is “vital to reach out to other disciplines that foregrounded how the socio-cultural notions of gender interact with language(s)” before we can discussing how gender inequality is encoded into machine translation systems (p. 847).

In a review of gender bias research in NLP, Sun et al. (2019) states in its future work section that “with few exceptions... work on debiasing NLP has assumed that the protected attribute being discriminated against is binary”(p. 8). This treatment of non-binary gender—relegation to the future work section, with only a couple sentences of acknowledgement—though, slowly changing, has long been the standard in most gender bias works, if non-binary gender is even acknowledged at all. As Devinney et al. (2022) identifies, the vast majority of research concerned with gender bias in NLP operationalizes gender

following the cisnormative ‘folk model’ (Keyes, 2018), in which gender is binary, immutable, and physiological. Popular benchmarks for measuring gender bias rely heavily on heteronormative views of gender, such as WinoBias (Zhao et al., 2018), which attempts to anonymize gender by swapping female and male entities in relation to different occupations. Many papers on gender bias do not define or theorize gender, instead conflating gender with sex, bodies, and pronouns, overlooking the experience of trans and genderqueer people. As established methodologies and datasets for evaluation continue to take on heteronormative frameworks, challenging the “classic” and “standard” binary approach will become even more difficult.

Research in bias surrounding sexual orientations, on the other hand, is much sparser and has mostly been limited to understanding hate speech. A binarity is still found in the work here. In HateCheck, a suite of tests to measure weak points in hate speech detection models, the target groups considered are women, gay people, black people, etc., isolating each axis of identity into a dominant and non-dominant group and ignoring the spectrum of queer identities (Röttger et al., 2021). In contrast, Dixon et al. (2018) produce a synthetic dataset where a larger collection of identity terms are slotted into templates of toxic (“I hate all queers”) and non-toxic phrases (“I am queer”), but this is to ignore the much more subtle representational biases present in language models. Stereotypes and implicit heteronormative biases have been largely overlooked, with only some recent developments.

Overall, a nuanced understanding of how gender and sexual identities interplay with language, as well as a closer interrogation with people’s lived experiences and how bias in systems are connected to real-world harm, is lacking from standard NLP bias research. Though significant progress has been made in terms of binary gender bias, it is important to set out a framework which can facilitate advancements in more complex understandings of identity.

5. Current queer bias research and recommendations

The following reviews some recent attempts at addressing queer bias in NLP systems, loosely separated into gender-inclusive bias research and broader queer bias research.

Cao and Daumé III (2020) develop two gender-inclusive datasets for analyzing how annotators introduce human bias into datasets and to test coreference resolution systems (e.g. resolving which pronouns refer to which entities) on non-binary and transgender people. Dev et al. (2021) examine the representational erasure of non-binary neopronouns xe and ze in current language tools, negative associations with gender identity words, and how language models explicitly misgender non-binary persons. Havens et al. (2022) presents a taxonomy of gendered biased language that incorporates trans and gender diverse identities as well as gender uncertainty, which can be used to create future bias datasets. Ovalle et al. (2023a) analyze popular large language models on TANGO, a dataset which documents examples of misgendering and harmful responses to gender disclosure.

Felkner et al. (2022) create a benchmark dataset focusing on gay and straight relationships and show that finetuning LLMs on a corpus of tweets from the LGBTQ+ community mitigates queer bias. Vásquez et al. (2022) create a corpus of tweets for studying heteronormative language focusing on binary gender terms. Barikeri et al. (2021) introduced RedditBias, a bias evaluation dataset created from Reddit comments which match target groups (lesbian, gay, bisexual, sapphic, etc.) with stereotypical attributes (mentally ill, flamboyant,

sinful, etc.). Lastly, Nozza et al. (2022) consider sentence completion of language models when the subjects belong to the LGBTQIA+ community.

From these works, I outline below open questions and recommendations towards a more unified framework to approaching queer bias.

Bias definitions and metrics: How do we define bias for queer experiences? What measures should we use to evaluate bias? Different NLP tasks require different notions of bias to appropriately capture how distributional skews may affect real-world outcomes. There are many considerations needed to understand which texts and outputs may be harmful, stereotypical, or denigrating, including the speaker of a text, the audience, the mode used to communicate, the social context a NLP system is used for. Although there will likely not be a one-size-fits-all approach to defining bias, an important first step is to ground bias research in social literature, following Cao and Daumé III (2020), Dev et al. (2021), and Vásquez et al. (2022) above. Doing so can establish a broader understanding of social and cultural norms in which a NLP system is situated, as well as more nuanced considerations of the interplay between language and social constructions and clearer definitions of the stakeholders and attributes involved. Most importantly, it is necessary that bias is designed with regard to potential for real harm. Following Ovalle et al. (2023a), bias definitions should be explicitly defined and informed directly from the lived experiences of and harms identified by queer stakeholders. It is necessary to continue to interface with current discourse and modern understandings of identity.

Data, ethics, and privacy: How can we ethically include queer identities in datasets for bias evaluation and mitigation? How can we build towards more affirming representations of historically and systematically excluded peoples? To ensure that queer people are protected rather than threatened by NLP systems, it is important to construct datasets which are representative of different experiences without creating undue burdens or risks and which always provide the option to opt out of participating. Research should consider using texts authored by and about queer communities, following Felkner et al. (2022) and Ovalle et al. (2023a), which will provide the most accurate representations of queer experiences. Whether datasets should include specific features such as identity groups is uncertain and may depend case by case. If they are included, it is important to consider how NLP systems may be misused to infer information or identify membership in a group, or how this may raise privacy concerns for the people included in the dataset. Otherwise, it may be helpful to link data to clearly defined harm cases. In addition, datasets should be constructed with the positionality of annotators in mind (Cao and Daumé III, 2020; Havens et al., 2022; Vasquez et al., 2022). Annotators may need additional situational context to not introduce their own biases into their labels—for example, improperly classifying a slur which has been reclaimed by the queer community.


Perspective and participatory methods: How do we reduce our own biases when conducting bias research? How do we ensure mitigating bias along one axis or definition may not unintentionally contribute to bias in unanticipated ways? It is difficult and perhaps impossible to predict every scenario in which an NLP system will be deployed and cause harm. Therefore, it is important for researchers to engage with their own positionality throughout the research process and involve queer participants in the design process. This positionality may arise for example, in prioritizing only one subgroup of the queer community (returning to the homosexual vs. heterosexual binary) or overlooking global and local contexts. Queer people should not be

treated as one homogenous group but rather a collection of communities and individuals who have varying experiences. Finally, as organizers of Queer in AI note, “Involving users as co-designers holds great potential for dismantling power relations and empowering marginalized communities that are disproportionately impacted by AI” (Ovalle et al., 2023b, p.1).

6. Conclusion

This paper sets out an argument for increased attention towards queer bias in NLP systems, beginning with an analysis of unique risks faced by queer communities and exploring the difficulties of analyzing queer bias. It then considers standard theories of gender and sexual orientation bias, as well as more recent works which have begun to tackle queer bias. Finally, it offers recommendations and points towards a framework for future work in queer bias. These include clearly defining harms and bias metrics, collecting data carefully and ethically, considering the positionality of authors and data annotators, and involving queer stakeholders in the research process. There is a promising increase in NLP conferences organized around fairness, community building of LGBTQ+ researchers in machine learning, and overall visibility of queer concerns in AI (Jethwani et al., 2022). However, with the speed at which NLP systems have begun to transform our day-to-day life, research in queer bias must be accelerated as well. Debiasing NLP is a critical step to more just futures for queer and all communities.

References

- Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2022). Machine bias. *In Ethics of data and analytics* (pp. 254-264). Auerbach Publications.
- Barikeri, S., Lauscher, A., Vulić, I., & Glavaš, G. (2021, August). RedditBias: A Real-World Resource for Bias Evaluation and Debiasing of Conversational Language Models. *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)*, 1941-1955. 10.18653/v1/2021.acl-long.151
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big? . In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623).
- Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020, July). Language (Technology) is Power: A Critical Survey of “Bias” in NLP. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5454-5476.
<https://aclanthology.org/2020.acl-main.485/>

- Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *Advances in neural information processing systems*, 29. <https://dl.acm.org/doi/10.5555/3157382.3157584>
- Burr, V., & Dick, P. (2017). Social Constructionism. In *The Palgrave Handbook of Critical Social Psychology*. Palgrave Macmillan. https://link.springer.com/chapter/10.1057/978-1-137-51018-1_4
- Caliskan, A. (2021, May). Detecting and mitigating bias in natural language processing. *The Brookings Institution's Artificial Intelligence and Emerging Technology (AIET) Initiative*. <https://www.brookings.edu/articles/detecting-and-mitigating-bias-in-natural-language-processing/>
- Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334), 183-186. <https://www.science.org/doi/10.1126/science.aal4230>
- Cameron, D. (2005, December). Language, Gender, and Sexuality: Current Issues and New Directions. *Applied Linguistics*, 26(4), 482-502. <https://doi.org/10.1093/applin/ami027>
- Cao, Y. T., & Daume III, H. (2020, July). Toward Gender-Inclusive Coreference Resolution. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4568–4595. 10.18653/v1/2020.acl-main.418
- Chen, Z., Zhang, J. M., Sarro, F., & Harman, M. (2023). A Comprehensive Empirical Study of Bias Mitigation Methods for Machine Learning Classifiers. *ACM Transactions on Software Engineering and Methodology*, 32(4), 1-30.
- Crawford, K. (2017). The Trouble with Bias. Keynote at Neural Information Processing Systems.
- Davis, L., & Reynolds, M. (2018, July). Gendered language and the educational gender gap. *Economic letters*, 168. <https://doi.org/10.1016/j.econlet.2018.04.006>
- Dennler, N., Ovalle, A., Singh, A., Soldaini, L., Subramonian, A., Tu, H., ... & Pinhal, J. D. J. D. P. (2023, August). Bound by the Bounty: Collaboratively Shaping Evaluation Processes for Queer AI Harms. In *Proceedings of the 2023 AAI/ACM Conference on AI, Ethics, and Society* (pp. 375-386).
- Dev, S., Monajatipoor, M., Ovalle, A., Subramonian, A., Phillips, J. M., & Chang, K.-W. (2021, November). Harms of Gender Exclusivity and Challenges in Non-Binary Representation in Language Technologies. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 1968–1994. <https://aclanthology.org/2021.emnlp-main.150.pdf>
- Devinney, H., Bjorklund, J., & Bjorklund, H. (2020, November). Crime and Relationship: Exploring Gender Bias in NLP Corpora. *SLTC 2020–The Eighth Swedish Language Technology Conference*. <https://umu.diva-portal.org/smash/get/diva2:1509712/FULLTEXT01.pdf>
- Devinney, H., Björklund, J., & Björklund, H. (2022, June). Theories of “Gender” in NLP Bias Research. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2083-2102. <https://doi.org/10.1145/3531146.3534627>
- Dixon, L., Li, J., Sorensen, J., Thain, N., & Vasserman, L. (2018, December). Measuring and Mitigating Unintended Bias in Text Classification. *Proceedings of the 2018 AAI/ACM Conference on AI, Ethics, and Society*, 67-73. <https://doi.org/10.1145/3278721.3278729>
- Dodge, J., Sap, M., Marasovic, A., Agnew, W., Ilharco, G., Groeneveld, D., Mitchell, M., & Gardner, M. (2021, November). Documenting Large Webtext Corpora: A Case Study on

- the Colossal Clean Crawled Corpus. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 1286-1305. 10.18653/v1/2021.emnlp-main.98
- Felkner, V. K., Chang, H.-C. H., Jang, E., & May, J. (2022). Towards WinoQueer: Developing a Benchmark for Anti-Queer Bias in Large Language Models. *arXiv preprint arXiv:2206.11484*.
- Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., & Roth, D. (2019, January). A comparative study of fairness-enhancing interventions in machine learning. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 329-338. <https://doi.org/10.1145/3287560.3287589>
- Geva, M., Goldberg, Y., & Berant, J. (2019, November). Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 1161–1166. <https://aclanthology.org/D19-1107.pdf>
- Gonen, H., & Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 609–614. <https://aclanthology.org/N19-1061.pdf>
- Hao, K. & Seetharaman, D. (2023, July 14). Cleaning Up ChatGPT Takes Heavy Toll on Human Workers. *Wall Street Journal*. <https://www.wsj.com/articles/chatgpt-openai-content-abusive-sexually-explicit-harassment-kenya-workers-on-human-workers-cf191483>
- Havens, L., Terras, M., Bach, B., & Alex, B. (2022, July). Uncertainty and Inclusivity in Gender Bias Annotation: An Annotation Taxonomy and Annotated Datasets of British English Text. *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*. 10.18653/v1/2022.gebnlp-1.4
- Hindes, S., & Fileborn, B. (2021). Reporting on sexual violence 'inside the closet': Masculinity, homosexuality and #MeToo. *Crime Media Culture*, 17(2), 163-184. <https://journals.sagepub.com/doi/10.1177/1741659020909872>
- Hovy, D., & Spruit, S. L. (2016, August). The Social Impact of Natural Language Processing. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, (pp. 591-598). <https://aclanthology.org/P16-2096/>
- Hovy, D., & Yang, D. (2021, June). The Importance of Modeling Social Factors of Language: Theory and Practice. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 588-602. <https://aclanthology.org/2021.naacl-main.49/>
- Introducing ChatGPT*. (2022, November 30). OpenAI. <https://openai.com/blog/chatgpt>
- Jethwani, H., Subramonian, A., Agnew, W., Bleil, M., Arora, S., Ryskina, M., & Xiong, J. (2022, July). Queer in AI. *XRDS: Crossroads, The ACM Magazine for Students*, 28(4), 18-21. <https://doi.org/10.1145/3538543>
- Jo, E. S., & Gebru, T. (2020, January). Lessons from archives: strategies for collecting sociocultural data in machine learning. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 306-316. <https://doi.org/10.1145/3351095.3372829>

- Käng, D. B. (2012, December). Kathoey “In Trend”: Emergent Genderscapes, National Anxieties and the Re-Signification of Male-Bodied Effeminacy in Thailand. *Asian Studies Review*, 475-494. <https://doi.org/10.1080/10357823.2012.741043>
- Kang, J., Xie, T., Wu, X., Maciejewski, R., & Tong, H. (2021). MultiFair: Multi-Group Fairness in Machine Learning. *arXiv preprint arXiv:2105.11069*. <https://www.semanticscholar.org/paper/MultiFair%3A-Multi-Group-Fairness-in-Machine-Learning-Kang-Xie/bab135b9fd0267e5319575e159dd2b28191b4b70>
- Keyes, O. (2018, November). The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. *Proceedings of the ACM on Human-Computer Interaction*, 1-22. <https://doi.org/10.1145/3274357>
- King, J. (2016). The violence of heteronormative language towards the queer community. *Aisthesis*, 7, 17-22. <https://pubs.lib.umn.edu/index.php/aisthesis/article/view/781>
- Magueresse, A., Carles, V., & Heetderk, E. (n.d.). Low-resource Languages: A Review of Past Work and Future Challenges. *arXiv preprint arXiv:2006.07264*. <https://arxiv.org/abs/2006.07264>
- Medina, C., Mahowald, L., Khattar, R., & Glass, A. (2022, June 1). *Fact Sheet: LGBT Workers in the Labor Market*. Center for American Progress. <https://www.americanprogress.org/article/fact-sheet-lgbt-workers-in-the-labor-market/>
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6), 1-35.
- Motschenbacher, H. (2021). Taking Queer Linguistics further: sociolinguistics and critical heteronormativity research. 149-179. <https://doi.org/10.1515/ijsl.2011.050>
- Nozza, D., Bianchi, F., Lauscher, A., & Hovy, D. (2022, May). Measuring Harmful Sentence Completion in Language Models for LGBTQIA+ Individuals. *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, 26-34. [10.18653/v1/2022.ltedi-1.4](https://arxiv.org/abs/2012.11865v1)
- Olivia, T. D., Antonialli, D. M., & Gomes, A. (2020, November). Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online. *Sexuality & Culture*, 25, 700-732. <https://doi.org/10.1007/s12119-020-09790-w>
- Ovalle, A., Goyal, P., Dhamala, J., Jagers, Z., Chang, K.-W., Galstyan, A., Zemel, R., & Gupta, R. (2023a). "I'm fully who I am": Towards Centering Transgender and Non-Binary Voices to Measure Biases in Open Language Generation. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. <https://doi.org/10.1145/3593013.3594078>
- Ovalle, A., Subramonian, A., Singh, A., Voelcker, C., Sutherland, D. J., ... & Stark, L. (2023b, June). Queer In AI: A Case Study in Community-Led Participatory AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1882-1895).
- Park, J. H., Shin, J., & Fung, P. (2018). Reducing Gender Bias in Abusive Language Detection. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2799-2804. <https://aclanthology.org/D18-1302/>
- Perrigo, B. (2023, January 18). Exclusive: OpenAI Used Kenyan Workers on Less Than \$2 Per Hour to Make ChatGPT Less Toxic *Time*. <https://time.com/6247678/openai-chatgpt-kenya-workers/>

- Pinchevski, A. (2023). Social media's canaries: content moderators between digital labor and mediated trauma. *Media, Culture & Society*, 45(1), 212-221.
- Prates, M. O. R., Avelar, P. H., & Lamb, L. C. (2020, May). Assessing gender bias in machine translation: a case study with Google Translate. *Neural Computing and Applications*, 32, 6363-6381. <https://doi.org/10.1007/s00521-019-04144-6>
- Queer in AI (2023) <https://www.queerinaai.com/>
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. *OpenAI blog*, 1(8), 9.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020, January). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21, 5485-5551. <https://jmlr.org/papers/volume21/20-074/20-074.pdf>
- Raji, I. D., Bender, E. M., Paullada, A., Denton, E., & Hanna, A. (2021). AI and the everything in the whole wide world benchmark. *arXiv preprint arXiv:2111.15366*.
- Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H., & Pierrehumbert, J. B. (2021, August). HATECHECK: Functional Tests for Hate Speech Detection Models. *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)*, 41-58. 10.18653/v1/2021.acl-long.4
- Savoldi, B., Gaido, M., Bentivogli, L., Negri, M., & Turchi, M. (2021). Gender Bias in Machine Translation. *Transactions of the Association for Computational Linguistics*, 9, 845-874. 10.1162/tacl_a_00401
- Shah, D., Schwartz, H. A., & Hovy, D. (2020, July). Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5248-5264. 10.18653/v1/2020.acl-main.468
- Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W., & Wang, W. Y. (2019, July). Mitigating Gender Bias in Natural Language Processing: Literature Review. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1630-1640. 10.18653/v1/P19-1159
- The Office of the Chief Statistician of the United States. (2023). *Recommendations on the Best Practices for the Collection of Sexual Orientation and Gender Identity Data on Federal Statistical Surveys*. The White House. <https://www.whitehouse.gov/wp-content/uploads/2023/01/SOGI-Best-Practices.pdf>
- Tomasev, N., McKee, K. R., Kay, J., & Mohamed, S. (2021, July). Fairness for Unobserved Characteristics: Insights from Technological Impacts on Queer Communities. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 254-265. <https://doi.org/10.1145/3461702.3462540>
- Vásquez, J., Bel-Enguix, G., Andersen, S. T., & Ojeda-Trueba, S.-L. (2022, July). HeteroCorpus: A Corpus for Heteronormative Language Detection. *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, 225-234. 10.18653/v1/2022.gebnlp-1.23
- Wang, A., Ramaswamy, V. V., & Russakovsky, O. (2022, June). Towards Intersectionality in Machine Learning: Including More Identities, Handling Underrepresentation, and Performing Evaluation. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 336-349. <https://doi.org/10.1145/3531146.3533101>

- Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., ... Gabriel, I. (2021, December). Ethical and social risks of harm from Language Models. *arXiv preprint arXiv:2112.04359*.
- Worthen, M. G.F. (2023, February). Queer identities in the 21st century: Reclamation and stigma. *Current Opinion in Psychology: Sexual & Gender Diversity in the 21st Century*, 49. <https://doi.org/10.1016/j.copsyc.2022.101512>
- Wu, D. (2023, February 7). *Twitch suspends GPT-3 'Seinfeld' parody after AI writes transphobic jokes*. The Washington Post. <https://www.washingtonpost.com/nation/2023/02/07/ai-seinfeld-transphobic-gpt3/>
- Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2017, September). Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2979–2989. 10.18653/v1/D17-1323
- Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2018, June). Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2*, 15-20. <https://aclanthology.org/N18-2003/>