Demographically Blind Models Can be Unfair: Fairness through Awareness

Anonymous ACL submission

Abstract

Debiasing methods for learning systems fall into two distinct philosophies of fairness: removing the use of protected attributes from the model, or including protected attributes in decision-making. However, the source of the bias that we seek to mitigate should dictate our choice of debiasing strategy. We categorize existing debiasing methods in these two fairness families, describe different types of biases, and show in controlled experiments that the choice of debiasing method should depend on the type of bias. Our results yield recommendations for practitioners moving forward.

1 Introduction

Numerous studies have demonstrated that NLP models can produce biased decisions and predictions through reliance on protected attributes (e.g. De-Arteaga et al., 2019). Models that screen for cancer, for example, may be less likely to suggest screenings for minority patients by identifying correlations between predictions and protected input attributes, like race, ethnicity, and gender, without critical social context. Without careful measurement and mitigation, trained models can perpetuate bias in tasks ranging from classification (Czarnowska et al., 2021; Zhang et al., 2020; Buolamwini and Gebru, 2018) to language generation (Blodgett et al., 2021; Cheng et al., 2023; Parrish et al., 2021; Dhamala et al., 2021). The exclusion of protected attributions can be insufficient; models identify other correlated attributes or embed demographic information into internal representations (Blodgett et al., 2016; Elazar and Goldberg, 2018).

Fairness through Unawareness (FTU) – exemplified in debiasing methods like adversarial debiasing (Zhang et al., 2020; Elazar and Goldberg, 2018; Han et al., 2021) or debiasing methods on embeddings (Liu et al., 2020; Huang et al., 2020) – aims to remove protected attributes to reduce their influence on learned representations and model

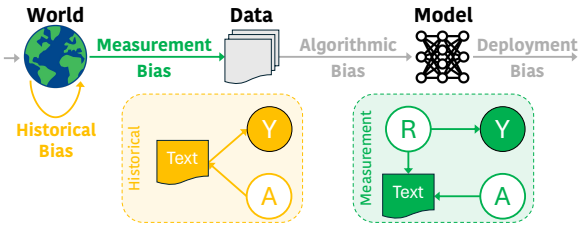

Figure 1: *Historical* and *Measurement Bias* on the text input, where Y is the predicted variable, A is protected attributes and R (not observed) is the real underlying feature that predicts Y.

attributes. FTU reflects one of two world views prevalent across decision-making in fields, such as finance, healthcare, politics, and education. This aligns with the philosophy of equality of *treatment*, where individuals should be treated equally despite differences in protected attributes.

Another view is equality of *outcomes*, where some differences in treatment may be needed for outcomes to be equal (Klarsfeld and Cachat-Rosset, 2021). This approach includes protected attributes in decision-making to ensure fair outcomes. This fairness distinction has been alluded to in prior work (Lipton et al., 2018; Barocas et al., 2019; Friedler et al., 2021; Hertweck et al., 2021), but here we propose to label these methods as Fairness through Awareness (FTA.) This paper connects these two world views (FTU, FTA) with current debiasing techniques in NLP and machine learning. Through theoretical and empirical analysis, we demonstrate that the choice of debiasing method crucially depends on the task and type of bias.

Complex factors lead to social biases that result in a complex interplay between protected attributes and decision (Friedler et al., 2021). We discuss two types of bias that illustrate a clear difference in the role of demographic attributes, *measurement bias* and *historical bias* (Baumann et al., 2023; Friedler et al., 2021); Fig. 1 shows these biases as well as others included for context (Suresh and Guttag,

2021). First, protected attributes can influence factors in a decision such that their inclusion promotes fair outcomes (*historical bias*: Zink et al., 2023). Alternatively, protected attributes can misinform a decision and promote surface correlations that lead to unfair outcomes (*measurement bias*: Zhang et al., 2020). We hypothesize that *measurement bias* requires FTU methods while *historical bias* necessitates FTA debiasing. We support this view by defining bias types and debiasing method families, as well as controlled synthetic experiments that allow us to change the type of bias present during training. Finally, we discuss the implications of these findings for debiasing language models.

2 Removing Decision Bias

While many NLP studies evaluate and mitigate social biases, few focus on identifying types of bias (Friedler et al., 2021; Hertweck et al., 2021). We utilize the taxonomy in Baumann et al. (2023), originally proposed by Suresh and Guttag (2021), and describe historical and measurement bias.

2.1 Types of Bias

Historical Bias arises when input features of the model and/or the target variable are influenced by some demographic attribute when they should not be, but have become part of the phenomenon to be captured by the machine learning system. An illustrative example is screening for cancer. An important factor in determining whether a patient should be screened for cancer is family history, where individuals with a history of cancer in their family have a higher likelihood of developing cancer themselves. Due to historical challenges in minority populations accessing healthcare in the United States, Black patients are less likely to possess accurate family history regarding cancer (Murff et al., 2005; Kupfer et al., 2006; Chavez-Yenter et al., 2022; Andoh, 2023). Models that build on positive correlations between family history of cancer and patient risk without considering racial confounds will bias against screenings for Black patients (Zink et al., 2023).

Historical bias is illustrated by the edge from the node "world" to itself in Fig. 1. More formally, given text input features *Text*, target variable Y and demographic variable A, the directed graph shows the historical bias of A on *Text*. The text features are part of the phenomenon resulting in Y and we typically assume we can measure these variables reliably. However, due to historical societal biases, the demographic variable A influences differences in text features and consequently in the observed target variable Y, even though, in principle, this should not be true.

Measurement Bias occurs when the variables causing the phenomenon are unobserved (R), and the observed text features are only proxies. As proxies, they are imperfect representations and may be influenced by, among other variables, demographic attributes A, as shown in Fig. 1 by the edge connecting *world* and *data*. An example of measurement bias can be found in kidney function measures, which guide physicians in choosing chemotherapy, nonprescription medication drugs, and anti-inflammatory drugs. Kidney function (R) often cannot be measured directly, therefore equations, such as the estimated glomerular filtration rate from serum creatinine (eCFRcr), are used instead as proxies. The eCFRcr uses race (A) as a feature because past work found that kidney function was different at similar levels of eCFRcr based on demographics, however, this association has been poorly justified and study replications remain inconclusive (Eneanya et al., 2019). Recently, studies have found that removing the race corrections in the eCFRcr leads to an increase in access to specialist care, kidney disease education, and kidney transplantation for African American patients (Diao et al., 2021).

More formally, given unobserved variable R, observed text features *Text*, label Y and demographic attributes A, Fig. 1 shows the directed graph portraying measurement bias on R through the input text features. The text features are not part of the phenomenon but are an imperfect proxy of the real phenomenon R, which has been influenced by A due to social factors.

2.2 Debiasing Methods

We introduce families of debiasing methods: Fairness through Unawareness (FTU) and Fairness through Awareness (FTA.) This distinction is present in conversations about fairness in other fields, as they are also known as *disparate treatment* (FTU) vs *disparate impact* (FTA) in economics and law (Lipton et al., 2018; Barocas et al., 2019), *We Are Equal* (FTU) vs *What You See Is What You Get* (FTA) (Friedler et al., 2021; Hertweck et al., 2021), and *race corrections* (FTA) in medicine (Zink et al., 2023).

Fairness through Awareness[1] methods seek to actively change model predictions based on protected attributes by either taking the demographic variables as input, applying a demographic-dependent training/regularizing loss, or modifying the prediction post-hoc. Traditionally, this could be achieved by including demographic attributes into the model input (Hovy, 2015), or adding a demographic-specific classification threshold during prediction (Hardt et al., 2016). In language models, these methods are less common, possibly because of the lack of demographic information available in datasets and because adding tabular data to language models is not trivial–requiring changes in model architecture or lower performance, e.g. (Suriyakumar et al., 2023). Examples of FTA methods for language models involve adding auxiliary losses during training/finetuning of models such as FairBatch (Roh et al., 2020; Foulds et al., 2020) and Disparate Learning Process (DLP) (Lipton et al., 2018), or, more recently, adding demographic features to prompts in few-shot learning (Röttger et al., 2021; Beck et al., 2024; Cheng et al., 2023; Deshpande et al., 2023; Aguirre et al., 2023; Santurkar et al., 2023).

Fairness through Unawareness methods reduce the influence of protected attributes on model prediction. While FTU in machine learning includes any method whose input does not explicitly include protected attributes, it is well understood that other features, such as text, can encode protected attributes in them (Elazar and Goldberg, 2018), and have been shown to use them as shortcuts that results in unfair behavior (Kotek et al., 2023). Further, it is often common to not include explicit demographic attributes in text, with a few exceptions (Cheng et al., 2023). We use FTU to include methods that seek to *actively* remove the influence of protected attributes on the input features, model parameters, or predictions. Some examples of these methods for language models are applied to the data directly in pre-processing (De-Arteaga et al., 2019), as adversarial debiasing for text classification (e.g. Zhang et al., 2020; Elazar and Goldberg, 2018; Beutel et al., 2017), for debiasing word embeddings (e.g. Bolukbasi et al., 2016; Caliskan et al., 2017; Chowdhury et al., 2021; Kaneko and Bollegala, 2021), iterative nullspace projection (INLP, Ravfogel et al., 2020; Subramanian et al., 2021; Ravfogel et al., 2022) and others (Chowdhury and Chaturvedi, 2022).

[1]The term Fairness through Awareness was the title of Dwork et al. (2012), and while the main contribution of the paper, a method later known as *individual fairness*, is an FTU method, an extension they present that includes the goal of "fair affirmative action" is considered FTA in our taxonomy.

2.3 Limitations of Fairness through Unawareness

What happens when we use FTU with models trained on biased data? Formally, assume that for a model $\bar{M}$ trained on dataset $D = \{x_i, y_i, a_i\}$, composed of input features $x_i \in X$, labels $y_i \in Y$, and demographic attributes $a_i \in A$, we observed that the predicted variable $\bar{M}(X) = \hat{Y}$ is somehow correlated with A. Thus, we train an unbiased model M with an FTU debiasing method.

Under measurement bias our observed features X are imperfect proxies influenced by demographic attributes A of the phenomenon R, where $Y = f_1(R)$, $X = f_2(R, A)$, where f_i are some naturally occurring function. Debiasing with FTU ensures that $\hat{y}_i$ are independent of a_i ($\hat{Y} \perp\!\!\!\perp A$). By eliminating the bias previously observed, while still allowing the model to approximate $f_1(R) \approx M(X)$, FTU can be effective for measurement bias.

Under historical bias X is accurately observed $R = X$, and $Y \perp\!\!\!\perp A|X$. The model trained with FTU, ensuring $\hat{Y} \perp\!\!\!\perp A$, looses important classification information as Y is dependent on A through X. Therefore, FTU is either not able to debias a model M or obtains a suboptimal model when historical bias is present.

3 Experiments

We now turn to a series of empirical demonstrations of the limitations of FTA versus FTU methods on datasets that contain *historical* and *measurement bias*, measuring the overall performance and fairness. We rely on synthetic datasets to control for the specific type of bias, which is not possible in natural datasets that arise from real social factors.

Data. We use a synthetic data generator[2] to create random variables (shown in Fig. 1) $R = -\beta_h^R A + N_R$, $N_R \sim \mathit{Gamma}(k_R, \theta_R)$, $A \sim \mathit{Ber}(p_A)$, and $P_R = R - \beta_m^R A + N_{P_R}$, $N_{P_R} \sim \mathcal{N}(0, \sigma^2_{P_R})$ as described in Baumann et al. (2023). Notably, β_h^R controls the presence and intensity of *historical bias* on input feature R, and β_m^R controls the *measurement bias* on the proxy feature P_R. We

[2]https://github.com/rcrupiISP/BiasOnDemand/tree/main

		Performance		Fairness	
	Model	μ	σ	μ	σ
Measurement Bias	Base	74.7	0.75	54.9	1.10
	FTU	73.0	1.43	57.3	3.86
	FTA	77.3	0.32	99.0	0.66
Historical Bias	Base	84.7	0.45	53.4	0.72
	FTU	83.4	0.60	53.4	1.24
	FTA	80.3	0.33	98.7	0.79

Table 1: Adversarial debiasing (FTU) results in fairer models without loss of performance for data with measurement bias, but worse and less fair performance for data with historical bias. Non-significant changes to the baseline ($p < .05$) over 20 random seeds in gray.

use their framework to create two datasets, one with historical bias on the input feature ($\beta_h^R = 3$) and the other with measurement bias on the proxy feature ($\beta_m^R = 3$), with the rest of variables with their default values. Appendix A.1 contains more details about the datasets as well as data statistics.

Models. We implement a standard neural network design with two feed-forward layers: an input layer and a classification layer (hidden size = 100). This model is labeled as *Base*. To represent FTU methods, we use the adversarial learning method proposed in Zhang et al. (2018)[3] and outlined in Appendix A.3. For the FTA methods, we use the fair threshold method (Threshold Optimizer)[4] initially proposed by Hardt et al. (2016), which chooses a different threshold for each demographic group based on a fairness constraint. The fairness constraint for both methods is demographic parity, which is also used to assess the fairness of the methods. Performance is measured in F1. We describe model architecture as well as more training details in Appendix A.2

Results. Table 1 shows the results for each method trained on both biased datasets. When the dataset includes measurement bias, both FTA and FTU methods yield models that are statistically more fair compared to the baseline, while maintaining similar or better overall performance. However, for historical bias, only the FTA method results in a fairer model than the baseline, while FTU performs worse and is less fair.

[3] https://fairlearn.org/main/user_guide/mitigation/adversarial.html

[4] https://fairlearn.org/v0.5.0/api_reference/fairlearn.postprocessing.html

4 Conclusion & Recommendations

Our field has embraced the importance of demographic fairness, developing many methods for debiasing trained models. However, many studies fail to differentiate between types of bias, nor identify which biases their methods are meant to combat (Blodgett et al., 2020). However, our analysis and experiments demonstrate that without these details, **debiasing methods can produce less fair models.** When *historical bias* is present and demographic attributes are important for prediction, FTU's objective directly conflicts with the prediction objective, resulting in a bias or a suboptimal predictor. On the other hand, there are also reasons why FTA methods are not effective and/or feasible: demographic variables may not be relevant to the task (De-Arteaga et al., 2019), demographic groups are too coarse to appropriately define harm towards people (Dwork et al., 2012), or it may be simply illegal for some tasks in the case of protected attributes. Failure to explicitly consider these factors will produce the opposite of the desired result.

We recommend the following best practices:

1. **Debate.** The effectiveness of debiasing methods depends on the bias type. Researchers should take a moment to reflect, debate, and decide what philosophy of debiasing method most applies for each task.
2. **Reporting for Researchers.** Researchers who develop debiasing methods should report on the type of method (FTA or FTU) and the assumed bias type(s) of interest.
3. **Reporting for Practitioners.** Practitioners applying existing debiasing methods to new tasks or settings should report on the known and assumed types of bias in the data (historical or measurement bias), and how the choice of debiasing methods addresses these biases. We caution against using debiasing methods without first understanding the source of bias.

Our community does not evaluate NLP systems in isolation; our choice of methods is meant to address specific social and data biases, whether due to historical factors, measurement issues, or other complex social issues. We must understand how our technical choices promote fairness of treatment or fairness of outcomes; without engaging in these conversations we cannot achieve our goals. We believe our analysis and recommendations will lead to more effective efforts to create fair NLP systems.

5 Ethical Considerations & Limitations

In this work we described two types of biases and how they interact with our classification of debiasing methods, however, we acknowledge there are many other types of bias, as shown in Figure 1, for which we did not explore the impact of the debiasing methods discussed here. The distinction we make, FTU vs FTA, may not have a significant difference in fairness or performance under other types of biases; however, this does not affect the scope of our claims and conclusions, as we found scenarios where debiasing methods can produce less fair models, thus affecting the choice of debiasing method families.

In addition, we assume that datasets have a single type of bias, however, it is likely that real life scenarios contain multiple types of biases at once, making the choice of FTU vs FTA harder to make from a theoretical point of view. Our experiments were performed under controlled settings in order to properly test our hypothesis where we ensured only one type of bias was introduced, however, the inter-relation of language and society is complex and is unlikely to produce datasets with only one type of social bias. This highlights the importance of our first recommendation, *debate*, as researchers will have to reflect and decide what philosophy to use in uncertain scenarios.

References

Carlos Aguirre, Kuleen Sasse, Isabel Cachola, and Mark Dredze. 2023. Selecting shots for demographic fairness in few-shot learning with large language models. *arXiv preprint arXiv:2311.08472*.

Joana E Andoh. 2023. The stories we don't know. *JAMA*, 329(18):1551–1551.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. Fairness and machine learning. fairmlbook. org.

Joachim Baumann, Alessandro Castelnovo, Riccardo Crupi, Nicole Inverardi, and Daniele Regoli. 2023. Bias on demand: A modelling framework that generates synthetic data with bias. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1002–1013.

Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2024. Sensitivity, performance, robustness: Deconstructing the effect of sociodemographic prompting. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2589–2615.

Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. 2017. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of african-american english. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130.

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 Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.

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

Daniel Chavez-Yenter, Melody S Goodman, Yuyu Chen, Xiangying Chu, Richard L Bradshaw, Rachelle Lorenz Chambers, Priscilla A Chan, Brianne M Daly, Michael Flynn, Amanda Gammon, et al. 2022. Association of disparities in family history and family cancer history in the electronic health record with sex, race, hispanic or latino ethnicity, and language preference in 2 large us health care systems. *JAMA network open*, 5(10):e2234574–e2234574.

Myra Cheng, Esin Durmus, and Dan Jurafsky. 2023. Marked personas: Using natural language prompts to measure stereotypes in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1504–1532.

Somnath Basu Roy Chowdhury and Snigdha Chaturvedi. 2022. Learning fair representations via rate-distortion maximization. *Transactions of the Association for Computational Linguistics*, 10:1159–1174.

Somnath Basu Roy Chowdhury, Sayan Ghosh, Yiyuan Li, Junier Oliva, Shashank Srivastava, and Snigdha Chaturvedi. 2021. Adversarial scrubbing of demographic information for text classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 550–562.

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying social biases in nlp: A generalization and empirical comparison of extrinsic fairness metrics. *Transactions of the Association for Computational Linguistics*, 9:1249–1267.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1236–1270.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 862–872.

James A Diao, Gloria J Wu, Herman A Taylor, John K Tucker, Neil R Powe, Isaac S Kohane, and Arjun K Manrai. 2021. Clinical implications of removing race from estimates of kidney function. *Jama*, 325(2):184–186.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21, Brussels, Belgium. Association for Computational Linguistics.

Nwamaka Denise Eneanya, Wei Yang, and Peter Philip Reese. 2019. Reconsidering the consequences of using race to estimate kidney function. *Jama*, 322(2):113–114.

James R Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. 2020. An intersectional definition of fairness. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 1918–1921. IEEE.

Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2021. The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making. *Communications of the ACM*, 64(4):136–143.

Xudong Han, Timothy Baldwin, and Trevor Cohn. 2021. Diverse adversaries for mitigating bias in training. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2760–2765, Online. Association for Computational Linguistics.

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Corinna Hertweck, Christoph Heitz, and Michele Loi. 2021. On the moral justification of statistical parity. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 747–757.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 752–762, Beijing, China. Association for Computational Linguistics.

Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. 2020. Reducing sentiment bias in language models via counterfactual evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 65–83, Online. Association for Computational Linguistics.

Masahiro Kaneko and Danushka Bollegala. 2021. Debiasing pre-trained contextualised embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1256–1266, Online. Association for Computational Linguistics.

Alain Klarsfeld and Gaëlle Cachat-Rosset. 2021. Equality of treatment, opportunity, and outcomes: mapping the law. In *Oxford Research Encyclopedia of Business and Management*.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, CI '23, page 12–24, New York, NY, USA. Association for Computing Machinery.

Sonia S Kupfer, Sarah McCaffrey, and Karen E Kim. 2006. Racial and gender disparities in hereditary colorectal cancer risk assessment: the role of family history. *Journal of Cancer Education*, 21.

Zachary Lipton, Julian McAuley, and Alexandra Chouldechova. 2018. Does mitigating ml's impact disparity require treatment disparity? *Advances in neural information processing systems*, 31.

Haochen Liu, Jamell Dacon, Wenqi Fan, Hui Liu, Zitao Liu, and Jiliang Tang. 2020. Does gender matter? towards fairness in dialogue systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4403–4416, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Harvey J Murff, Daniel Byrne, Jennifer S Haas, Ann Louise Puopolo, and Troyen A Brennan. 2005. Race and family history assessment for breast cancer. *Journal of general internal medicine*, 20(1):75–80.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R Bowman. 2021. Bbq: A hand-built bias benchmark for question answering. *arXiv preprint arXiv:2110.08193*.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. *arXiv preprint arXiv:2004.07667*.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan D Cotterell. 2022. Linear adversarial concept erasure. In *International Conference on Machine Learning*, pages 18400–18421. PMLR.

Yuji Roh, Kangwook Lee, Steven Euijong Whang, and Changho Suh. 2020. Fairbatch: Batch selection for model fairness. *arXiv preprint arXiv:2012.01696*.

Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet B Pierrehumbert. 2021. Two contrasting data annotation paradigms for subjective nlp tasks. *arXiv preprint arXiv:2112.07475*.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR.

Shivashankar Subramanian, Xudong Han, Timothy Baldwin, Trevor Cohn, and Lea Frermann. 2021. Evaluating debiasing techniques for intersectional biases. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2492–2498.

Harini Suresh and John Guttag. 2021. A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and access in algorithms, mechanisms, and optimization*, pages 1–9.

Vinith Menon Suriyakumar, Marzyeh Ghassemi, and Berk Ustun. 2023. When personalization harms performance: reconsidering the use of group attributes in prediction. In *International Conference on Machine Learning*, pages 33209–33228. PMLR.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340.

Haoran Zhang, Amy X Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. 2020. Hurtful words: quantifying biases in clinical contextual word embeddings. In *proceedings of the ACM Conference on Health, Inference, and Learning*, pages 110–120.

A Zink, Z Obermeyer, and E Pierson. 2023. Race corrections in clinical models: Examining family history and cancer risk.

A Experiment Details

In this section we provide more details about the implementation and use of the dataset and models for the experiments in §3.

A.1 Data Details

We use a synthetic data generator[5] to create random variables (shown in Fig. 1) as described in Baumann et al. (2023):

$$
\begin{aligned}
R &= -\beta_h^R A + N_R \\
N_R &\sim \textit{Gamma}(k_R, \theta_R) \\
A &\sim \textit{Ber}(p_A) \\
Q &\sim \textit{Bin}(K, p_Q(R, A)) \\
P_R &= R - \beta_m^R A + N_{P_R} \\
N_{P_R} &\sim \mathcal{N}(0, \sigma_{P_R}^2) \\
S &= \alpha_R R - \alpha_Q Q - \beta_h^Y A + N_S \\
N_S &\sim \mathcal{N}(0, \sigma_S^2) \\
Y &= \mathbf{1}_{\{S > \bar{P}_S\}}
\end{aligned}
$$

Here, R is a random variable drawn from a Gamma distribution, that optionally depends on β_h^R which controls the presence and intensity of *historical bias* from A on the feature. A is a binary random variable drawn from a Bernoulli distribution. P_R is the proxy variable that may be optionally influenced by β_m^R which controls the *measurement bias*.

We use their framework to create two datasets, one with *historical bias*, where R (along with Q) are the input features $X = [R, Q]$ with $\beta_h^R = 3$; and the other with *measurement bias*, where P_R (along with Q) are the input features $X = [P_R, Q]$ with $\beta_m^R = 3$. The rest of the variables are left with their default values. Each dataset contains 100K data points, with a train-test split (.66-.33) with stratified sampling on the demographics ensuring each split has equal demographic distributions. Table 2 shows the statistics for both datasets and their train-test splits.

A.2 Models

We implement a standard neural network design on PyTorch[6] with two fully connected feed-forward layers: an input layer (`input_size` = 2, `output_size` = `hidden_dim` = 100) and a classification layer (`input_size` = `hidden_dim` = 100, `output_size` = 2) with a ReLu activition function in between. This model was trained using a binary cross-entropy loss, we used Adam as the optimizer with a `learning rate` = .001 and batch size = 32 for one epoch of the dataset. This model is labeled as *Base*.

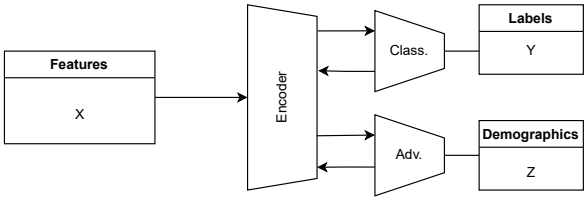

Figure 2: Schematic of adversarial debiasing method for an encoder based model.

To represent FTU methods, we use the adversarial learning method proposed in Zhang et al. (2018) and outlined in Appendix A.3. We use the FairLearn implementation (`AdversarialFairnessClassifier`[7]) with both the task classifier and the adversarial classifier with the same architecture as *Base*, the $\alpha = .7$, and demographic parity as the fairness constraint.

For the FTA methods, we use the fair threshold method (`ThresholdOptimizer`[8]) initially proposed by Hardt et al. (2016), which chooses a different threshold for each demographic group based on a fairness constraint. The classifier has the same architecture as the *Base*, and the fairness constraint is also demographic parity.

We measure performance in F1, and fairness in demographic parity. Table 1 show the average and standard deviation of both performance and fairness of the models trained over 20 random seeds. Statistical significance was calculated by random samples of the train data and training each model with different random seeds to obtain a distribution of test scores. Then, we perform an ANOVA test with subsequent pairwise t-tests with Bonferroni corrections.

A.3 Adversarial Debiasing

Briefly, adversarial debiasing involves adding an adversarial loss to the main objective during training with the goal to discourage the model's hidden

[5] https://github.com/rcrupiISP/BiasOnDemand/tree/main

[6] https://pytorch.org/

[7] https://fairlearn.org/main/user_guide/mitigation/adversarial.html

[8] https://fairlearn.org/v0.5.0/api_reference/fairlearn.postprocessing.html

	Historical Bias								*Measurement Bias*							
	train				test				train				test			
	R	Q	A	Y	R	Q	A	Y	P_R	Q	A	Y	P_R	Q	A	Y
count	67k	67k	67k	67k	33k	33k	33k	33k	67k	67k	67k	67k	33k	33k	33k	33k
μ	4.51	1.50	0.50	0.45	4.51	1.50	0.50	0.45	4.51	1.50	0.50	0.44	4.51	1.50	0.50	0.44
σ	4.52	0.86	0.50	0.50	4.47	0.87	0.50	0.50	4.94	0.86	0.50	0.50	4.90	0.87	0.50	0.50
min	-2.99	0.00	0.00	0.00	-2.96	0.00	0.00	0.00	-8.93	0.00	0.00	0.00	-9.22	0.00	0.00	0.00
25%	1.36	1.00	0.00	0.00	1.37	1.00	0.00	0.00	1.05	1.00	0.00	0.00	1.12	1.00	0.00	0.00
50%	3.70	1.00	1.00	0.00	3.74	1.00	1.00	0.00	3.87	1.00	1.00	0.00	3.89	1.00	1.00	0.00
75%	6.81	2.00	1.00	1.00	6.86	2.00	1.00	1.00	7.24	2.00	1.00	1.00	7.23	2.00	1.00	1.00
max	41.41	3.00	1.00	1.00	42.32	3.00	1.00	1.00	47.15	3.00	1.00	1.00	42.99	3.00	1.00	1.00

Table 2: Data statistics for the train and test split for both datasets.

representations from predicting demographic attributes. In a typical neural network style model, this is implemented by adding a demographic attribute classification layer and incorporate its loss to the main classification layer's loss. Fig. 2 shows a schematic of an adversarial model.

More formally, we assume data points $x_i \in X$ with paired target variables $y_i \in Y$, and $z_i \in Z$ are the corresponding demographic attribute for $\{x_i, y_i\}$. We train a model $M(X)$ that is composed of: an encoder $f(X)$, that takes as input the features x_i and outputs hidden representations h_i, as well as a classification layer $c(H)$ that takes as input the hidden representations h_i and outputs the prediction $\hat{y}$. Adversarial debiasing seeks hidden representations h_i that are independent of z_i. This is achieved if there is not a demographic classifier $adv(H)$ that predicts the attributes z_i from h_i. Let θ be the parameters of the model, $\theta = \{\theta_f, \theta_c, \theta_{adv}\}$. To such an end, the training procedure concretely seeks to optimize both objectives jointly:

$$\min_{\theta} M(X, \theta) \triangleq \mathcal{L}(X, Y, [\theta_f \cup \theta_c]) - \mathcal{L}_{adv}(X, Z, [\theta_f \cup \theta_{adv}])$$

Importantly, this method assumes that Y can be predicted without information from Z, the demographic attributes. Otherwise, the adversarial loss would be in direct contradiction with the classifier loss, and would obtain a suboptimal classifier.