Are You Man Enough? Even Fair Algorithms Conform to Societal Norms

Anonymous Authors[1]

Abstract

We introduce Societal Norm Bias (SNoB), a subtle but consequential type of discrimination that may be exhibited by machine learning classification algorithms, even when these systems achieve group fairness objectives. This work illuminates the gap between definitions of algorithmic group fairness and concerns of harm based on adherence to societal norms. We study this issue through the lens of gender bias in occupation classification from online biographies. We quantify SNoB by measuring how an algorithm's predictions are associated with gender norms. This framework reveals that for classification tasks related to male-dominated occupations, fairness-aware classifiers favor biographies whose language aligns with masculine gender norms. We compare SNoB across fairness intervention techniques, finding that post-processing interventions do not mitigate this bias at all.

1. Introduction

As automated decision-making systems play a growing role in our daily lives, concerns about algorithmic unfairness have come to light (Buolamwini & Gebru, 2018; Noble, 2018; Stark et al., 2020). To avoid algorithmic discrimination based on sensitive attributes, various approaches to measure and achieve fairness have been proposed. These approaches are typically based on *group fairness*, which partitions a population into groups based on a protected attribute (e.g. gender, race, religion) and then aims to equalize some metric of the system across the groups.

Group fairness makes the implicit assumption that a group is defined solely by the possession of particular characteristics (Hu & Kohler-Hausmann, 2020), ignoring the heterogeneity within groups. It does not account for the complex, multidimensional nature of concepts like gender and race (Hanna et al., 2020; Butler, 2011), thus overlooking the various axes along which bias may occur, such as an individual's adherence to societal norms.

We characterize Societal Norm Bias (SNoB)—the associations between an algorithm's predictions and individuals' adherence to societal norms—as a source of algorithmic unfairness. We study SNoB through the task of occupation classification on a dataset of online biographies. In this setting, masculine/feminine SNoB occurs when an algorithm favors biographies written in ways that adhere to masculine/feminine gender norms, respectively. We examine how existing fairness intervention techniques, based on categorical gender labels, neglect this issue. Discrimination based on gender norms has implications of concrete harms, which are documented in the social science literature (Section 2.3).

Our approach measures how an algorithm's predictions are associated with masculine or feminine gender norms based on natural language features. This framework quantifies an algorithm's bias on another dimension of gender beyond explicit binary labels. Using this framework to evaluate fairness interventions, we analyze the differences among how these approaches encode gender norms. We find that approaches that improve group fairness still exhibit SNoB. In particular, post-processing approaches are most closely aligned to gender norms. These associations may lead to representational and allocational harms for feminine-expressing people in male-dominated occupations (Bartl et al., 2020; Blodgett et al., 2020). Furthermore, when fairness-aware algorithms exhibit SNoB, these harms are not only perpetuated but also obscured by claims of group fairness.

2. Background

2.1. The Multiplicity of Gender

The term "gender" is used as a proxy for different ideas depending on the context (Keyes et al., 2021). It may mean *gender identity*, which is one's "felt, desired or intended identity" (Glick et al., 2018), or *gender expression*, which is how one "publicly expresses or presents their gender... others perceive a person's gender through these attributes" (Commission). These concepts are also related to *gender norms*, i.e. "the standards and expectations to which women and men generally conform," including personality traits, behaviors, occupations, and appearance (Agius & Tobler, 2012). These various notions of social gender encompass much more than the categorical gender labels that are used as the basis for group fairness approaches (Cao & III, 2019). We focus on discrimination related to the ways that individuals' gender expression adhere to societal gender norms.

2.2. Gender Bias in Automated Hiring

Audit studies reveal that employers tend to discriminate against women (Bertrand & Mullainathan, 2004; Johnson et al., 2016). These biases are also replicated in automated hiring. For example, previous work measures the gender gap in error rates of an occupation classification algorithm (De-Arteaga et al., 2019). Many in academia and industry alike have been motivated to mitigate these concerns (Raghavan et al., 2020; Bogen & Rieke, 2018; Sánchez-Monedero et al., 2020). LinkedIn developed a post-processing approach for ranking candidates so that their candidate recommendations are demographically representative of the broader candidate pool; their system is deployed across a service affecting more than 600 million users (Geyik et al., 2019). Other intervention techniques have also been proposed (Dwork et al., 2018; Romanov et al., 2019). These approaches share a reliance on categorical gender labels to measure fairness.

2.3. Harms Related to Gender Norms in the Workplace

Our concerns about the use of gender norms in machine learning systems are grounded in studies of how gender norms have been operationalized in various occupations, causing harm to gender minorities. It is well-established that "occupations are socially and culturally 'gendered'" (Stark et al., 2020); many jobs in science, technology, and engineering are perceived as masculine (Ensmenger, 2015; Light, 1999). Women in these fields have been found to perform their gender in particular ways to gain respect and acceptance from their peers, in turn fostering a "masculine" environment that is hostile to women (Powell et al., 2009).

In social psychology, *descriptive stereotypes* are attributes believed to characterize women as a group. Heilman (2001; 2012) study how the perceived lack of fit between feminine stereotypic attributes and male gender-typed jobs result in gender bias and impede women's careers.

When these patterns are replicated by SNoB in machine learning algorithms, this results in two types of harms. The associations that we highlight may lead to 1) representational harm, when actual members of the occupation are made invisible, and 2) allocational harm, when certain individuals are allocated fewer career opportunities based on their gender (Bartl et al., 2020; Blodgett et al., 2020).

3. Methods

To study SNoB, we focus on the use of gender norms in occupation classification, a component of automated recruiting. We assume that a "fair" occupation classification algorithm should not exhibit gender bias, including SNoB, since someone's career potential is not related to their gender. There ought to be no association between the classifier's predictions and any concept of gender (pronouns, expression, etc.). However, unlike for gender pronouns, there is no ground-truth label for other notions of gender in our biography dataset. Thus, we use a data-driven approach to measure each biography's adherence to gender norms. We validate this approach by comparing it to crowd-sourced notions of gender norms. We then compare the degree to which individuals' adherence to gender norms is correlated with occupation classification predictions. We introduce metrics to quantify masculine and feminine SNoB in occupation classifiers on two different scales: single occupation association and cross-occupation association.

3.1. Occupation Classification

3.1.1. Dataset

We use the dataset[1] and task described by De-Arteaga et al. (2019). The dataset, containing 397,340 biographies spanning twenty-eight occupations, is obtained by filtering the Common Crawl for online biographies.

Each biography is labeled with its gender based on the use of "she" or "he" pronouns; biographies that contain neither pronoun are excluded. (In Appendix C, we study a small set of biographies with nonbinary pronouns.) Let H_c, S_c be the sets of biographies in occupation c using "he" and "she" respectively. $|H_c|, |S_c|$ are the numbers of biographies in the respective sets. To preserve the ratios between $|H_c|$ and $|S_c|$, we use a stratified split to create the training, validation, and test datasets, containing 65%, 10%, and 25% of the biographies respectively. We use the data to train and evaluate an algorithm that predicts a biography's occupation title from the subsequent sentences.

3.1.2. Semantic Representations

For the occupation classification algorithm, we use three semantic representations with different degrees of complexity: bag-of-words, word embeddings, and BERT. In the bag-of-words (BOW) representation, a biography b is represented as a sparse vector of the frequencies of the words in b. BOW is widely used in settings where interpretability is important. In the word embedding (WE) representation, b is represented by an average of the fastText word embeddings (Bojanowski et al., 2017; Mikolov et al., 2018) for the words in b. Previous work demonstrates that the WE representation captures semantic information effectively (Adi et al., 2016). For the BOW and WE representations, we train a one-versus-all logistic regression model with L_2 regularization on the training set, as done by De-Arteaga et al. (2019). The BERT contextualized word embedding model (Devlin et al., 2018) is state-of-the-art for various natural language processing tasks, and it has been widely adopted for many uses. Unlike the other language representations,

[1] The dataset is publicly available at http://aka.ms/biasbios.

a biography's encoding is context-dependent. We fine-tune the BERT model, which pre-trains deep bidirectional representations from unlabeled English text (Wolf et al., 2020), for the occupation classification task.

3.2. Quantifying Gender Norms

We leverage the natural language properties of our biography dataset to measure how much a biography's gender expression aligns with societal norms. There are many differences in the ways that language is used to describe people of different genders (Menegatti & Rubini, 2017), and in the ways that people of different genders choose to use language (Argamon et al., 2003). Gender also affects the ways that people are perceived (Madera et al., 2009). See Appendix A for details.

One brute-force way to measure biographies' adherence to gender norms is to obtain crowdsourced gender ratings for every word used in the dataset, and then score each biography using these ratings. Because human-labeled corpora of gendered words (Crawford et al., 2004; Cryan et al., 2020) are limited to a few hundred words, while our biography dataset has tens of thousands of unique words, we take a machine learning approach to quantify these notions rather than relying on the human-labeled corpora alone. We train a classifier G on the biographies dataset to distinguish between whether a biography is labeled with "she" or "he." For an individual biography, we use G's predicted probability of "s/he" as a measure of how much the biography aligns with feminine/masculine gender norms.

To validate that G learns a meaningful notion of gender norms, we compare its similarity to human-labeled gender scores. Specifically, for a corpus of 600 words with gender scores labeled via crowdsourcing (Crawford et al., 2004), we compare G's weights of these words to the human-labeled gender scores reported in the study. We find a strong correlation (Pearson's r-value 0.76) between these values. See Appendix Figure 3 for details. Also, in Appendix E, we perform a robustness test using a gender classifier that omits occupation-relevant words.

3.3. Measuring Masculine and Feminine SNoB

For a given biography b, the occupation classification algorithm Y_c outputs the probability $Y_c(b)$ that the individual belongs to occupation c. The gender classifier G outputs the probability $G(b)$ that the individual's biography is labeled with "she". To evaluate SNoB, we use the correlation r_c between $G(b)$ and $Y_c(b)$, the predicted probabilities from the two classifiers, across the "she" bios in the occupation. Specifically, we compute Pearson's correlation coefficient r_c between $\{Y_c(b)|b \in S_c\}$ and $\{G(b)|b \in S_c\}$. The magnitude of r_c is a measure of the degree of SNoB exhibited by the occupation classifier. A positive/negative value indicates that more feminine/masculine language is rewarded by Y_c.

Consider $p_c = \frac{|S_c|}{|S_c|+|H_c|}$, i.e. the fraction of biographies in occupation c that use "she." If $p_c < 0.5$, c is male-dominated, and vice versa. We find that r_c is more negative in more male-dominated occupations, i.e. individuals whose biographies are more aligned to masculine gender norms are also more likely to be correctly predicted by the occupation classification algorithm. Since r_c is computed from S_c, these associations are present *within* the gender group. Thus, classification for male-dominated occupations algorithms operationalize gendered language, privileging not only the referential gender of pronouns but also the "she" biographies with more masculine words and writing styles.

We observe a trend between r_c and p_c: in more gender-imbalanced occupations, r_c is larger in magnitude. Let $\mathbf{r}_C = \{r_c|c \in C\}$, $\mathbf{p}_C = \{p_c|c \in C\} \in [0,1]^{|C|}$, where C is the set of occupations. The covariance $\text{Cov}(\mathbf{p}_c, \mathbf{r}_c)$ quantifies this trend for an algorithmic approach across all the occupations. We use covariance rather than correlation because the latter does not capture the range of the values, i.e. the magnitude of the slopes in Figures 1 and 2, while for an individual classifier, we use correlation r_c since the range is less important than the relative rankings across the individuals in S_c.

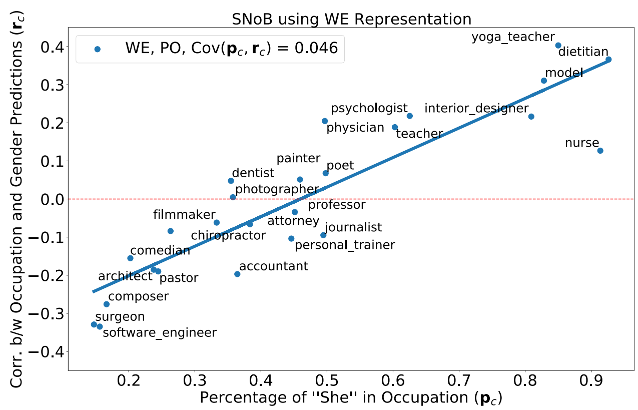

Figure 1. **SNoB Across Occupations.** The extent to which an algorithm's predictions align with gender norms (y-axis) is correlated with the gender imbalance in the occupation (x-axis). Ideally, without any SNoB, the correlation $r_c = 0$, so every point would lie on the dotted red line. Other representations (BOW, BERT) have similar trends. Note that these values are the same for the fairness-unaware approach as the post-processing approach.

4. Analysis of Fairness Approaches

We evaluate two paradigms of algorithmic group fairness approaches, post-processing and in-processing techniques. These approaches are based on the goal of mitigating Gap^{RMS}, the group fairness metric used by Romanov et al. (2019) and De-Arteaga et al. (2019); see Appendix B for details. We present these approaches and compare their SNoB.

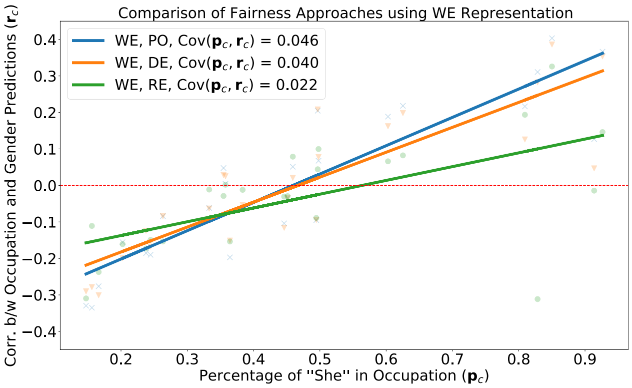

Figure 2. **Comparing fairness interventions.** While SNoB persists across group fairness interventions, it is somewhat mitigated by the in-processing approaches; the slopes of their best-fit lines, which correspond to $\text{Cov}(\mathbf{p}_c, \mathbf{r}_c)$, are smaller than that of PO.

4.1. Fairness Intervention Techniques

Post-processing (PO) fairness approaches apply an intervention after training the algorithm to balance some metric across groups (Pleiss et al., 2017; Kamiran et al., 2012; Lohia et al., 2019; Hardt et al., 2016). PO is relatively cost-effective and has been deployed in large-scale automated recruiting systems (Geyik et al., 2019). Since PO techniques do not change the ordering within a group, r_c remains identical to that of the approach without any fairness intervention (Figure 1). Thus, the interventions may continue to privilege individuals who align with the occupation's gender norms.

We also consider various *in-processing* group fairness approaches, which modify the algorithm at training time. In the decoupled (DE) approach, a separate classifier is trained for each groups (Dwork et al., 2018). In the reductions approach (RE), a classification task is reduced to a sequence of cost-sensitive classification problems (Agarwal et al., 2018). RE is the primary in-processing mitigation method in the Fairlearn Python package (Bird et al., 2020). Covariance Constrained Loss (CoCL) adds a constraint to the loss function that minimizes the covariance between an individual's predicted probability and the word embedding of their name. Romanov et al. (2019) validate CoCL's effectiveness in reducing Gap^{RMS} on the same biographies dataset.

4.2. Comparing Approaches

We use $\text{Cov}(\mathbf{p}_c, \mathbf{r}_c)$ (Section 3.3) to compare the associations for different fairness approaches (Figure 2, Table 1). The PO approach has the largest value of $\text{Cov}(\mathbf{p}_c, \mathbf{r}_c)$, i.e. the strongest associations with gender norms. For PO, the predicted probabilities and within-group ranking of the individuals are unchanged from the fairness-unaware occupation classification algorithm. Even when the desired statistical metric is perfectly met, i.e. $\text{Gap}^{\text{RMS}} = 0$, these correlations remain. For PO, the group fairness and SNoB metrics seem to be unrelated; the mitigation of one is not informative about the presence of the other. The in-processing approaches (DE, RE, CoCL) mitigate this observed association since their $\text{Cov}(\mathbf{p}_c, \mathbf{r}_c)$ are lower compared to that of PO. However, $\text{Cov}(\mathbf{p}_c, \mathbf{r}_c)$ remains nonzero, which suggests that gender norms continue to be leveraged in these approaches (see Appendix D for more analysis on the mechanisms). Since in-processing approaches are typically more expensive to implement than PO, there are trade-offs between ease of implementation, classifier accuracy, and association with gender norms. Unlike in PO, there are more complex relationships between Gap^{RMS} and $\text{Cov}(\mathbf{p}_c, \mathbf{r}_c)$ for in-processing approaches. Both Gap^{RMS} and $\text{Cov}(\mathbf{p}_c, \mathbf{r}_c)$ are larger for WE, DE than WE, RE (Table 1). While BOW, CoCL has larger $\text{Cov}(\mathbf{p}_c, \mathbf{r}_c)$ than BOW, DE approaches, its Gap^{RMS} is smaller. This suggests that there is no straightforward correspondence between Gap^{RMS} and $\text{Cov}(\mathbf{p}_c, \mathbf{r}_c)$.

Table 1. Although post-processing (PO) fairness intervention techniques mitigate Gap^{RMS} the most, they have higher values of $\text{Cov}(\mathbf{p}_c, \mathbf{r}_c)$ compared to in-processing approaches. This suggests that the latter are more effective at reducing SNoB than PO. For BOW and WE, the one-versus-all Y_c accuracy is averaged across all occupations. For BERT, the model is a multi-class classifier.

Approach	Y_c Accuracy	Gap^{RMS}	$\text{Cov}(\mathbf{p}_c, \mathbf{r}_c)$
BOW, PO	0.95	0	**0.023**
BOW, DE	0.96	0.10	0.014
BOW, CoCL	0.96	0.086	0.021
WE, PO	0.97	0	**0.046**
WE, DE	0.94	0.060	0.040
WE, RE	0.88	0.035	0.022
BERT, PO	0.85	0	**0.021**
BERT, DE	0.85	0.22	**0.021**

5. Future Work

We measure associations between algorithmic predictions and gender norms in occupation classification, revealing that SNoB is the strongest in post-processing approaches. Since occupation classification is a subtask of automated recruiting, the associations may have significant consequences in people's lives.

More broadly, we characterize how algorithms may discriminate based on SNoB, a non-categorical aspect of a sensitive attribute. By illuminating the axes along which discrimination may occur, our work sets the stage for progress in mitigating these harms. We hope to explore algorithmic approaches to reducing these associations as well as socio-technical considerations of how the intersectionality (Crenshaw, 1990) between different dimensions of a sensitive attribute affects an algorithm.

References

Adi, Y., Kermany, E., Belinkov, Y., Lavi, O., and Goldberg, Y. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *arXiv preprint arXiv:1608.04207*, 2016.

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. A reductions approach to fair classification. In *International Conference on Machine Learning*, pp. 60–69. PMLR, 2018.

Agius, S. and Tobler, C. *Trans and intersex people. Discrimination on the grounds of sex, gender identity and gender expression*. Office for Official Publications of the European Union, 2012.

Argamon, S., Koppel, M., Fine, J., and Shimoni, A. R. Gender, genre, and writing style in formal written texts. *Text - Interdisciplinary Journal for the Study of Discourse*, 23(3), 2003. doi: 10.1515/text.2003.014.

Bartl, M., Nissim, M., and Gatt, A. Unmasking contextual stereotypes: Measuring and mitigating bert's gender bias. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pp. 1–16, 2020.

Bertrand, M. and Mullainathan, S. Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American economic review*, 94(4):991–1013, 2004.

Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., and Walker, K. Fairlearn: A toolkit for assessing and improving fairness in AI. Technical Report MSR-TR-2020-32, Microsoft, May 2020. URL https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/.

Blodgett, S. L., Barocas, S., Daumé III, H., and Wallach, H. Language (technology) is power: A critical survey of "bias" in nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5454–5476, 2020.

Bogen, M. and Rieke, A. Help wanted. *An Examination of Hiring Algorithms, Equity, and Bias*, 2018.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5: 135–146, 2017.

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

Butler, J. *Gender trouble: Feminism and the subversion of identity*. routledge, 2011.

Cao, Y. T. and III, H. D. Toward gender-inclusive coreference resolution. *CoRR*, abs/1910.13913, 2019. URL http://arxiv.org/abs/1910.13913.

Commission, O. H. R. Gender identity and gender expression. URL http://www.ohrc.on.ca/en/policy-preventing-discrimination-because-gender-identity-and-gender-expression/3-gender-identity-and-gender-expression.

Crawford, J. T., Leynes, P. A., Mayhorn, C. B., and Bink, M. L. Champagne, beer, or coffee? a corpus of gender-related and neutral words. *Behavior Research Methods, Instruments, & Computers*, 36(3):444–458, 2004. doi: 10.3758/bf03195592.

Crenshaw, K. Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stan. L. Rev.*, 43:1241, 1990.

Cryan, J., Tang, S., Zhang, X., Metzger, M., Zheng, H., and Zhao, B. Y. Detecting gender stereotypes: Lexicon vs. supervised learning methods. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–11, 2020.

De-Arteaga, M., Romanov, A., Wallach, H., Chayes, J., Borgs, C., Chouldechova, A., Geyik, S., Kenthapadi, K., and Kalai, A. T. 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*, pp. 120–128, 2019.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Dwork, C., Immorlica, N., Kalai, A. T., and Leiserson, M. Decoupled classifiers for group-fair and efficient machine learning. In *Conference on Fairness, Accountability and Transparency*, pp. 119–133. PMLR, 2018.

Ensmenger, N. "beards, sandals, and other signs of rugged individualism": masculine culture within the computing professions. *Osiris*, 30(1):38–65, 2015.

Geyik, S. C., Ambler, S., and Kenthapadi, K. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2221–2231, 2019.

Glick, J. L., Theall, K., Andrinopoulos, K., and Kendall, C. For data's sake: dilemmas in the measurement of gender minorities. *Culture, health & sexuality*, 20(12): 1362–1377, 2018.

Hanna, A., Denton, E., Smart, A., and Smith-Loud, J. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 501–512, 2020.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 3323–3331, 2016.

Heilman, M. E. Description and prescription: How gender stereotypes prevent women's ascent up the organizational ladder. *Journal of social issues*, 57(4):657–674, 2001.

Heilman, M. E. Gender stereotypes and workplace bias. *Research in organizational Behavior*, 32:113–135, 2012.

Hu, L. and Kohler-Hausmann, I. What's sex got to do with machine learning? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 513–513, 2020.

Johnson, S. K., Hekman, D. R., and Chan, E. T. If there's only one woman in your candidate pool, there's statistically no chance she'll be hired. *Harvard Business Review*, 26(04), 2016.

Kamiran, F., Karim, A., and Zhang, X. Decision theory for discrimination-aware classification. In *2012 IEEE 12th International Conference on Data Mining*, pp. 924–929. IEEE, 2012.

Keyes, O., May, C., and Carrell, A. You keep using that word: Ways of thinking about gender in computing research. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–23, 2021.

Light, J. S. When computers were women. *Technology and culture*, 40(3):455–483, 1999.

Lohia, P. K., Ramamurthy, K. N., Bhide, M., Saha, D., Varshney, K. R., and Puri, R. Bias mitigation post-processing for individual and group fairness. In *Icassp 2019-2019 ieee international conference on acoustics, speech and signal processing (icassp)*, pp. 2847–2851. IEEE, 2019.

Madera, J. M., Hebl, M. R., and Martin, R. C. Gender and letters of recommendation for academia: agentic and communal differences. *Journal of Applied Psychology*, 94(6):1591, 2009.

Menegatti, M. and Rubini, M. Gender bias and sexism in language. In *Oxford Research Encyclopedia of Communication*. 2017.

Mikolov, T., Grave, É., Bojanowski, P., Puhrsch, C., and Joulin, A. Advances in pre-training distributed word representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

Moon, R. From gorgeous to grumpy: adjectives, age and gender. *Gender & Language*, 8(1), 2014.

Noble, S. U. *Algorithms of oppression: How search engines reinforce racism*. nyu Press, 2018.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. On fairness and calibration. *arXiv preprint arXiv:1709.02012*, 2017.

Powell, A., Bagilhole, B., and Dainty, A. How women engineers do and undo gender: Consequences for gender equality. *Gender, work & organization*, 16(4):411–428, 2009.

Raghavan, M., Barocas, S., Kleinberg, J., and Levy, K. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 469–481, 2020.

Romanov, A., De-Arteaga, M., Wallach, H., Chayes, J., Borgs, C., Chouldechova, A., Geyik, S., Kenthapadi, K., Rumshisky, A., and Kalai, A. T. What's in a name? reducing bias in bios without access to protected attributes. *arXiv preprint arXiv:1904.05233*, 2019.

Sánchez-Monedero, J., Dencik, L., and Edwards, L. What does it mean to'solve'the problem of discrimination in hiring? social, technical and legal perspectives from the uk on automated hiring systems. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 458–468, 2020.

Snyder, K. The resume gap: Are different gender styles contributing to tech's dismal diversity. *Fortune Magazine*, 2015.

Stark, L., Stanhaus, A., and Anthony, D. L. "i don't want someone to watch me while im working": Gendered views of facial recognition technology in workplace surveillance. *Journal of the Association for Information Science and Technology*, 71(9):1074–1088, 2020. doi: 10.1002/asi.24342.

Wagner, C., Garcia, D., Jadidi, M., and Strohmaier, M. It's a man's wikipedia? assessing gender inequality in an online encyclopedia. In *Proceedings of the International

AAAI Conference on Web and Social Media, volume 9, 2015.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.