An Intersectional Definition of Fairness

James R. Foulds, Rashidul Islam, Kamrun Naher Keya, Shimei Pan
Department of Information Systems
University of Maryland, Baltimore County, USA
{jfoulds, islam.rashidul, kkeya1, shimei}@umbc.edu

A full version of this paper with proofs and further results is available at: https://arxiv.org/abs/1807.08362**.**

***Abstract*—We propose differential fairness, a multi-attribute definition of fairness in machine learning which is informed by intersectionality, a critical lens arising from the humanities literature, leveraging connections between differential privacy and legal notions of fairness. We show that our criterion behaves sensibly for any subset of the set of protected attributes, and we prove economic, privacy, and generalization guarantees. We provide a learning algorithm which respects our differential fairness criterion. Experiments on the COMPAS criminal recidivism dataset and census data demonstrate the utility of our methods.**

***Index Terms*—fairness in AI, AI and society, 80% rule, privacy**

I. Introduction and Motivation

The increasing impact of artificial intelligence and machine learning technologies on many facets of life, from commonplace movie recommendations to consequential criminal justice sentencing decisions, has prompted concerns that these systems may behave in an unfair or discriminatory manner [2], [19]. A number of studies have subsequently demonstrated that bias and fairness issues in AI are both harmful and pervasive [1], [4], [5]. The AI community has responded by developing a broad array of mathematical formulations of fairness and learning algorithms which aim to satisfy them [3], [10], [13], [20], [24]. Fairness, however, is not a purely technical construct, having social, political, philosophical and legal facets [6]. The necessity has now become clear for interdisciplinary analyses of fairness in AI and its relationship to society, to civil rights, and to the social goals which are to be achieved by mathematical fairness definitions, which have not always been made explicit [18]. In this work, we address the specific challenges of fairness in AI that are motivated by **intersectionality**, an analytical lens from the third-wave feminist movement which emphasizes that civil rights and feminism should be considered simultaneously rather than separately [9]. We propose an **intersectional AI fairness criterion** and perform a theoretical analysis of its properties relating to diverse fields including the **humanities**, **law**, **privacy**, **economics**, and **statistical machine learning**.

This work was performed under the following financial assistance award: 60NANB18D227 from U.S. Department of Commerce, National Institute of Standards and Technology. This material is based upon work supported by the National Science Foundation under Grant No. IIS 1850023. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

The principle of *intersectionality* emphasizes that systems of oppression built into society lead to *systematic disadvantages along intersecting dimensions*, which include not only gender, but also race, nationality, sexual orientation, disability status, and socioeconomic class [7]–[9], [14], [17], [22]. These systems are interlocking in their effects on individuals at *each intersection of the affected dimensions*. Intersectionality thus implies the use of multiple *protected attributes*, and has further implications. Many AI fairness definitions aim (implicitly or otherwise) to uphold the principle of *infra-marginality*, which states that differences between protected groups in the distributions of "merit" or "risk" (e.g. the probability of carrying contraband at a policy stop) should be taken into account when determining whether bias has occurred [21]. In short, the *infra-marginality* principle makes the implicit assumption that society is a fair, level playing field, and thus differences in "merit" or "risk" between groups in data and predictive algorithms are often to be considered legitimate. In contrast, *intersectionality* theory posits that these **distributions of merit and risk are often influenced by unfair societal processes**. In ideal intersectional fairness, since ability to succeed is affected by unfair processes, it is desired that this unfairness is corrected and individuals achieve their true potential [23]. Assuming individuals' unbiased potential does not substantially differ across protected groups, this implies that parity between groups, and intersectional subgroups, is typically desirable.[1]

In the machine learning literature, the previous AI fairness definition most relevant to intersectionality is *statistical parity subgroup fairness (SF)* [15]. We adapt the notation of [16] to all definitions in this paper. Suppose $M(\mathbf{x})$ is a (possibly randomized) mechanism which takes an instance $\mathbf{x} \in \chi$ and produces an outcome $y \in \mathcal{Y}$ for the corresponding individual, $S_1, \ldots, S_p$ are discrete-valued protected attributes, $A = S_1 \times S_2 \times \ldots \times S_p$, and θ is the distribution which generates $\mathbf{x}$. Each individuals' data $\mathbf{x}_i$ is stored in a dataset D on a secure server. The mechanism $M(\mathbf{x})$ could, for example, be a deep learning model for a lending decision, A could be the applicant's possible gender and race, and θ the joint distribution of credit scores and protected attributes. The protected attributes are included in the attribute vector

[1] Disparity could still be desirable if there are legitimate confounders which depend on protected groups, e.g. choice of department that individuals apply to in college admissions. We address this in the extended arXiv paper.

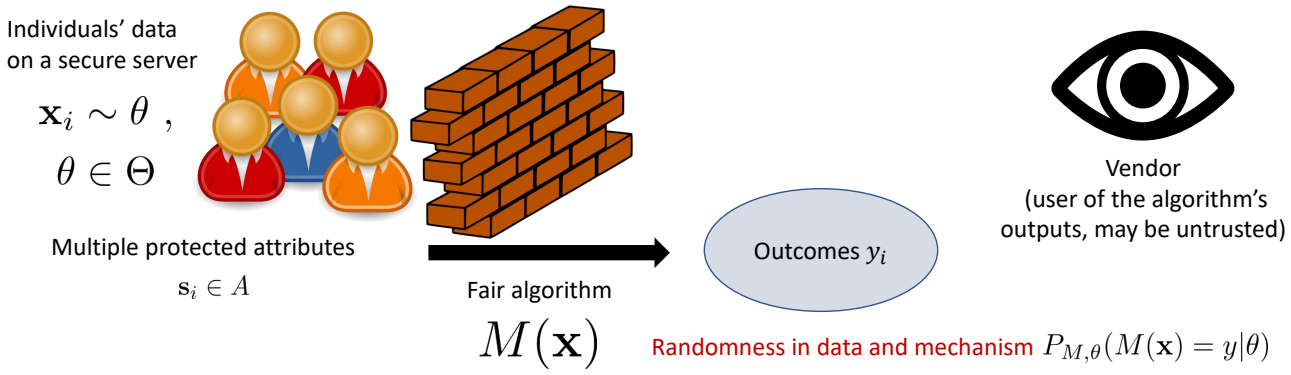


Fig. 1. Diagram of the setting for the proposed differential fairness criterion.

$\mathbf{x}$, although $M(\mathbf{x})$ is free to disregard them (e.g. if this is disallowed). The setting is illustrated in Figure 1.

Definition I.1. *(Statistical Parity Subgroup Fairness [15]) Let $\mathcal{G}$ be a collection of protected group indicators $g : A \to \{0, 1\}$, where $g(\mathbf{s}) = 1$ designates that an individual with protected attributes $\mathbf{s}$ is in group g. Assume that the classification mechanism $M(\mathbf{x})$ is binary, i.e. $y \in \{0, 1\}$.*

Then $M(\mathbf{x})$ is γ-statistical parity subgroup fair *with respect to θ and $\mathcal{G}$ if for every $g \in \mathcal{G}$,*

$$|P_{M,\theta}(M(\mathbf{x}) = 1) - P_{M,\theta}(M(\mathbf{x}) = 1|g(\mathbf{s}) = 1)| \times P_\theta(g(\mathbf{s}) = 1) \leq \gamma \ . \quad (1)$$

From an intersectional perspective, one concern with SF is that it does not protect minority groups, often marginalized by society, and whose protection intersectionality emphasizes. The term $P_\theta(g(\mathbf{s}) = 1)$ weights the "per-group (un)fairness" for each group g, i.e. Equation 1 applied to g alone, by its proportion of the population, thereby *downweighting the consideration of minorities*.

II. Differential Fairness (DF) Measure

We propose an alternative fairness criterion which is more concordant with intersectionality, including its treatment of minorities and its other provable theoretical properties. We first motivate our criterion from a legal perspective. Consider the 80% rule, established in the Code of Federal Regulations [12] as a guideline for establishing disparate impact in violation of anti-discrimination laws such as Title VII of the Civil Rights Act of 1964. The 80% rule states that there is legal evidence of adverse impact if the ratio of probabilities of a particular favorable outcome, taken between a disadvantaged and an advantaged group, is less than 0.8:

$$P(M(\mathbf{x}) = 1|\text{group A})/P(M(\mathbf{x}) = 1|\text{group B}) < 0.8 \ . \quad (2)$$

Our proposed criterion, which we call **differential fairness (DF)**, extends the 80% rule to protect multi-dimensional intersectional categories, with respect to multiple output values. We similarly restrict ratios of outcome probabilities between groups, but instead of using a predetermined fairness threshold at 80%, we measure fairness on a sliding scale that can be interpreted similarly to that of *differential privacy*, a definition of privacy for data-driven algorithms [11]. Differential fairness measures the **fairness cost** of mechanism $M(\mathbf{x})$ with a parameter ϵ.

Definition II.1. *A mechanism $M(\mathbf{x})$ is* ϵ-differentially fair (DF) *with respect to (A, Θ) if for all $\theta \in \Theta$ with $\mathbf{x} \sim \theta$, and $y \in Range(M)$,*

$$e^{-\epsilon} \leq \frac{P_{M,\theta}(M(\mathbf{x}) = y|\mathbf{s}_i, \theta)}{P_{M,\theta}(M(\mathbf{x}) = y|\mathbf{s}_j, \theta)} \leq e^\epsilon \ , \quad (3)$$

for all $(\mathbf{s}_i, \mathbf{s}_j) \in A \times A$ where $P(\mathbf{s}_i|\theta) > 0$, $P(\mathbf{s}_j|\theta) > 0$.

In Equation 3, $\mathbf{s}_i, \mathbf{s}_j \in A$ are tuples of *all* protected attribute values, e.g. gender, race, and nationality, and Θ is a set of distributions θ which could plausibly generate each instance $\mathbf{x}$.[2] For example, Θ could be the set of Gaussian distributions over credit scores per value of the protected attributes, with mean and standard deviation in a certain range.

This is an intuitive **intersectional definition of fairness**: *regardless of the combination of protected attributes, the probabilities of the outcomes will be similar*, as measured by the ratios versus other possible values of those variables, for small values of ϵ. For example, the probability of being given a loan would be similar regardless of a protected group's intersecting combination of gender, race, and nationality, marginalizing over the remaining attributes in $\mathbf{x}$. If the probabilities are always equal, then $\epsilon = 0$, otherwise $\epsilon > 0$. We have arrived at our criterion based on the 80% rule, but it can also be derived as a special case of *pufferfish* [16], a generalization of differential privacy [11] which uses a variation of Equation 3 to hide the values of an arbitrary set of secrets.

III. Estimating Differential Fairness from Data

If $P_{M,\theta}$ is unknown, it can be estimated using the empirical distribution, or via a probabilistic model of the data. Assuming

[2]The possibility of multiple $\theta \in \Theta$ is valuable from a privacy perspective, where Θ is the set of *possible beliefs* that an adversary may have about the data, and is motivated by the work of [16]. We will however typically assume a single distribution, $\Theta = \{\theta\}$. Continuous protected attributes are also possible, in which case sums are replaced by integrals in our proofs.

discrete outcomes, $P_{Data}(y|\mathbf{s}) = \frac{N_{y,\mathbf{s}}}{N_{\mathbf{s}}}$, where $N_{y,\mathbf{s}}$ and $N_{\mathbf{s}}$ are empirical counts of their subscripted values in the dataset D. **Empirical differential fairness (EDF)** corresponds to verifying that for any y, $\mathbf{s}_i$, $\mathbf{s}_j$, we have

$$e^{-\epsilon} \leq \frac{N_{y,\mathbf{s}_i}}{N_{\mathbf{s}_i}} \frac{N_{\mathbf{s}_j}}{N_{y,\mathbf{s}_j}} \leq e^{\epsilon} \ . \tag{4}$$

We can adapt DF to measure fairness in data, i.e. outcomes assigned by a black-box algorithm or social process, by using (a model of) the data's generative process as the mechanism.

Definition III.1. *A labeled dataset* $D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$ *is* ϵ-differentially fair (DF) *in* A *with respect to model* $P_{Model}(\mathbf{x}, y)$ *if mechanism* $M(\mathbf{x}) = y \sim P_{Model}(y|\mathbf{x})$ *is* ϵ-differentially fair *with respect to* $(A, \{P_{Model}(\mathbf{x})\})$*, for* P_{Model} *trained on the dataset.*

In the long paper on the arXiv, we consider extensions of DF to handle confounder variables, and to measure the amplification of bias due to an algorithm.

IV. Properties of Differential Fairness (Proofs Given in the arXiv Paper)

Intersectionality: Differential fairness explicitly encodes protection of intersectional groups. For DF, we prove that this automatically implies fairness for *each of the protected attributes individually*, and indeed, *any subset* of the protected attributes. For example, by ensuring fairness at the intersection of gender, race, and nationality under our criterion, we also ensure the same degree of fairness between genders overall, and between gender/nationality pairs overall, and so on.

Theorem IV.1. ***(Intersectionality Property)*** *Let* M *be an* ϵ-differentially fair *mechanism in* (A, Θ)*,* $A = S_1 \times S_2 \times \ldots \times S_p$*, and let* $D = S_a \times \ldots \times S_k$ *be the Cartesian product of a nonempty proper subset of the protected attributes included in* A*. Then* M *is* ϵ-differentially fair *in* (D, Θ)*.*

Privacy: The ϵ-DF definition, and the resulting level of fairness obtained at any particular measured fairness parameter ϵ, can be interpreted by viewing the definition through the lens of privacy. Differential fairness ensures that given the outcome, *an untrusted vendor/adversary can learn very little about the protected attributes of the individual*, relative to their prior beliefs, assuming their prior beliefs are in Θ:

$$e^{-\epsilon} \frac{P(\mathbf{s}_i|\theta)}{P(\mathbf{s}_j|\theta)} \leq \frac{P(\mathbf{s}_i|M(\mathbf{x}) = y, \theta)}{P(\mathbf{s}_j|M(\mathbf{x}) = y, \theta)} \leq e^{\epsilon} \frac{P(\mathbf{s}_i|\theta)}{P(\mathbf{s}_j|\theta)} \ . \tag{5}$$

The privacy guarantee only holds if $\theta \in \Theta$, which may not always be the case. Regardless, the value of ϵ may typically be interpreted as a privacy guarantee against a "reasonable adversary."

Utility: An ϵ-differentially fair mechanism admits a disparity in expected utility of as much as a factor of $\exp(\epsilon) \approx 1 + \epsilon$ (for small values of ϵ) between pairs of protected groups with $\mathbf{s}_i \in A$, $\mathbf{s}_j \in A$, for any utility function. The proof follows the case of differential privacy [11], see the arXiv paper.

Generalization: To ensure that an algorithm is truly fair, the fairness properties obtained on a dataset must extend to the underlying population. We prove a generalization guarantee for estimating ϵ-DF although it is weaker than for subgroup fairness [15] – the price of protecting minority subgroups:

Theorem IV.2. ***(Generalization Property)*** *Fix a class of functions* $\mathcal{H}$*, which without loss of generality aim to discriminate the outcome* $y = 1$ *from any other value, denoted here as* $y = 0$*. For any conditional distribution* $P(y, \mathbf{x}|\mathbf{s})$ *given a group* $\mathbf{s}$*, let* $S \sim P^m$ *be a dataset consisting of* m *examples* $(\mathbf{x}_i, y_i)$ *sampled i.i.d. from* $P(y, \mathbf{x}|\mathbf{s})$*. Then for any* $0 < \delta < 1$*, with probability* $1 - \delta$*, for every* $h \in \mathcal{H}$*, we have:*

$$|P(y = 1|\mathbf{s}, h) - P_S(y = 1|\mathbf{s}, h)| \leq \tilde{O}\Big(\sqrt{\frac{VCDIM(\mathcal{H}) \log m + \log(1/\delta)}{m}}\Big) \ . \tag{6}$$

V. Learning Algorithm

Our learning algorithm uses the fairness cost as a regularizer to balance the trade-off between fairness and accuracy. We minimize, with respect to the classifier $M_{\mathbf{W}}(\mathbf{x})$'s parameters $\mathbf{W}$, a loss function $L_{\mathbf{X}}(\mathbf{W})$ plus a penalty on unfairness which is weighted by a tuning parameter $\lambda > 0$. We train fair neural networks using adaptive gradient descent (Adam) on our objective via backpropagation and automatic differentiation (*DF-Classifier*), and similarly for subgroup fairness (*SF-Classifier*). The learning objective for training data $\mathbf{X}$ becomes:

$$\min_{\mathbf{W}} [L_{\mathbf{X}}(\mathbf{W}) + \lambda R_{\mathbf{X}}(\epsilon)] \tag{7}$$

where $R_{\mathbf{X}}(\epsilon) = max(0, \epsilon_{M_{\mathbf{W}}(\mathbf{x})} - \epsilon_1)$ represents the fairness penalty term, and $\epsilon_{M_{\mathbf{W}}(\mathbf{x})}$ is the ϵ for $M_{\mathbf{W}}(\mathbf{x})$. If ϵ_1 is 0, this penalizes ϵ-DF, and if ϵ_1 is the data's ϵ, this penalizes the *bias amplification* by the algorithm (see the arXiv paper). To make the objective differentiable, $\epsilon_{M_{\mathbf{W}}(\mathbf{x})}$ is estimated using soft counts $P(y|\mathbf{x})$ from the classifier. Below, α is a Dirichlet smoothing parameter, and $N_{\mathbf{s}}$ is the count for group $\mathbf{s}$.

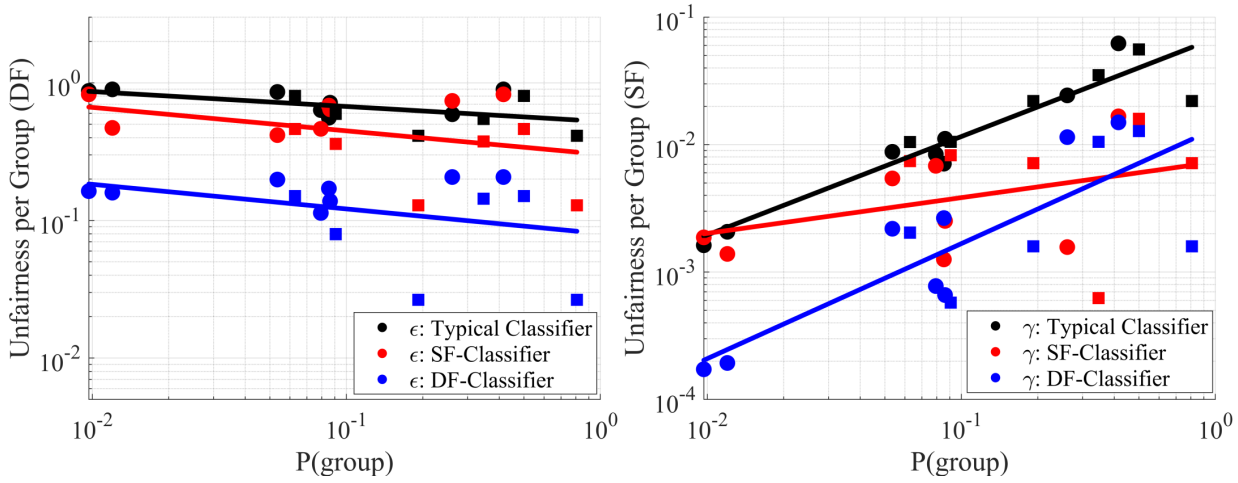
$$e^{-\epsilon} \leq \frac{\sum_{\mathbf{x} \in D: A = \mathbf{s}_i} P(y|\mathbf{x}) + \alpha}{N_{\mathbf{s}_i} + |\mathcal{Y}|\alpha} \frac{N_{\mathbf{s}_j} + |\mathcal{Y}|\alpha}{\sum_{\mathbf{x} \in D: A = \mathbf{s}_j} P(y|\mathbf{x}) + \alpha} \leq e^{\epsilon}$$

VI. Experiments and Conclusion

We performed experiments on the COMPAS dataset regarding a system that is used to predict criminal recidivism [1] (protected attributes: *race* and *gender*). Further experiments were performed on the Adult 1994 U.S. census income data from the UCI repository (protected attributes: *race*, *gender*, USA vs non-USA *nationality*), see the arXiv paper.[3]

An important goal of this work was to consider the impact of the fairness methods on minority groups. In Figure 2, we report the "per-group unfairness," defined as Equations 1 and 3 with one group held fixed, versus the group's probability (i.e. size) on the COMPAS dataset. Both methods improve their corresponding per-group unfairness measures over the typical classifier. On the other hand, the γ-SF metric only

[3]Predicted income, used for consequential decisions like housing approval, may result in *digital redlining* [2].



(a) Improvement in DF measures

(b) Improvement in SF measures

Fig. 2. Per-group measurements of (a) ϵ-DF and (b) γ-SF of the classifiers vs group size (probability), COMPAS dataset, calculated using Equations 1 and 3 with the group held fixed. Circles: intersectional subgroups. Squares: top-level groups. The methods improve fairness, both per group and overall, but SF-Classifier is seen to ignore minority groups in the overall γ-SF measurement, calculated as a worst-case over all groups.

assigns high per-group unfairness values to large groups in its measurement, so **minority groups are not able to influence the overall γ-SF unfairness**. This was not the case **for ϵ-DF metric, where groups of various sizes had similarly high per-group ϵ values**. Furthermore, the **DF-Classifier improved the per-group fairness under both metrics for groups of all sizes, while the SF-classifier did not improve the per-group γ-SF for small groups**. Further experiments, given in the arXiv paper, show that DF-Classifier and SF-Classifier behave similarly in terms of accuracy, and that they can be tuned to improve fairness with little loss in accuracy. Our overall conclusion is that the *DF-Classifier is able to achieve intersectionally fair classification with minor loss in performance, while providing greater protection to minority groups than when enforcing subgroup fairness*.

Acknowledgment

We thank Rosie Kar for valuable advice and feedback regarding intersectional feminism.

References

[1] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica, May*, 23, 2016.

[2] S. Barocas and A.D. Selbst. Big data's disparate impact. *Cal. L. Rev.*, 104:671, 2016.

[3] R. Berk, H. Heidari, S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, S. Neel, and A. Roth. A convex framework for fair regression. *FAT/ML Workshop*, 2017.

[4] T. Bolukbasi, K.-W. Chang, J.Y. Zou, V. Saligrama, and A.T. Kalai. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in NeurIPS*, 2016.

[5] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *FAT**, 2018.

[6] A. Campolo, M. Sanfilippo, M. Whittaker, A. Selbst K. Crawford, and S. Barocas. *AI Now 2017 Symposium Report*. AI Now, 2017.

[7] P.H. Collins. *Black feminist thought: Knowledge, consciousness, and the politics of empowerment (2nd ed.)*. Routledge, 2002 [1990].

[8] Combahee River Collective. A black feminist statement. In Z. Eisenstein, editor, *Capitalist Patriarchy and the Case for Socialist Feminism*. Monthly Review Press, New York, 1978.

[9] K. Crenshaw. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *U. Chi. Legal F.*, pages 139–167, 1989.

[10] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of ITCS*, pages 214–226, 2012.

[11] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Th. of Cryptography*, pages 265–284, 2006.

[12] Equal Employment Opportunity Commission. Guidelines on employee selection procedures. *C.F.R.*, 29.1607, 1978.

[13] M. Hardt, E. Price, N. Srebro, et al. Equality of opportunity in supervised learning. In *Advances in NeurIPS*, pages 3315–3323, 2016.

[14] b. hooks. *Ain't I a Woman: Black Women and Feminism*. South End Press, 1981.

[15] M. Kearns, S. Neel, A. Roth, and Z.S. Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *Proc. of ICML, PMLR 80*, pages 2569–2577, 2018.

[16] D. Kifer and A. Machanavajjhala. Pufferfish: A framework for mathematical privacy definitions. *TODS*, 39(1):3, 2014.

[17] A. Lorde. Age, race, class, and sex: Women redefining difference. In *Sister Outsider*, pages 114–124. Ten Speed Press, 1984.

[18] S. Mitchell, E. Potash, and S. Barocas. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv preprint arXiv:1811.07867*, 2018.

[19] S.U. Noble. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press, 2018.

[20] B. Salimi, L. Rodriguez, B. Howe, and D. Suciu. Interventional fairness: Causal database repair for algorithmic fairness. In *Proceedings of the 2019 Int. Conf. on Management of Data*, pages 793–810, 2019.

[21] C. Simoiu, S. Corbett-Davies, S. Goel, et al. The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics*, 11(3):1193–1216, 2017.

[22] S. Truth. Ain't I a woman?, 1851. Speech delivered at Women's Rights Convention, Akron, Ohio.

[23] C. Verschelden. *Bandwidth Recovery: Helping Students Reclaim Cognitive Resources Lost to Poverty, Racism, and Social Marginalization*. Stylus, 2017.

[24] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of EMNLP*, 2017.