Towards A Unifying Human-Centered AI Fairness Framework

Munshi Mahbubur Rahman
mrahman4@umbc.edu
University of Maryland, Baltimore County
Baltimore, Maryland, USA

Shimei Pan
shimei@umbc.edu
University of Maryland, Baltimore County
Baltimore, Maryland, USA

James R. Foulds
jfoulds@umbc.edu
University of Maryland, Baltimore County
Baltimore, Maryland, USA

ABSTRACT

Achieving fairness in AI systems is a critical yet challenging task due to conflicting metrics and their underlying societal assumptions, e.g., the extent to which racist and sexist societal processes are presumed to cause harm and the extent to which we should apply affirmative corrections. Moreover, these measures often contradict each other and might also make the AI system less accurate. This work takes a step towards a unifying human-centered fairness framework to guide stakeholders in navigating these complexities, including their potential incompatibility and the corresponding trade-offs. Our framework acknowledges the spectrum of fairness definitions —individual vs. group fairness, infra-marginal (politically conservative) vs. intersectional (politically progressive) treatment of disparities— allowing stakeholders to prioritize desired outcomes by assigning weights to various fairness considerations, trading them off against each other, as well as predictive performance, supporting stakeholders in exploring the impacts of their fairness choices to achieve a consensus solution. Our learning algorithms then ensure the resulting AI system reflects the stakeholder-chosen priorities. By enabling multi-stakeholder compromises, our framework can potentially mitigate individual analysts' subjectivity. We performed experiments to validate our methods on the UCI Adult census dataset and the COMPAS criminal recidivism dataset.

CCS CONCEPTS

• **Applied computing** → **Law, social and behavioral sciences**; • **Computing methodologies** → **Artificial intelligence**; **Machine learning**; • **Social and professional topics** → **User characteristics**.

KEYWORDS

Fairness in AI, AI & Society, Human-Centered Fairness, Fairness/ Performance Trade-offs.

ACM Reference Format:
Munshi Mahbubur Rahman, Shimei Pan, and James R. Foulds. 2024. Towards A Unifying Human-Centered AI Fairness Framework. In *International Conference on Information Technology for Social Good (GoodIT '24), September 04–06, 2024, Bremen, Germany*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3677525.3678645

GoodIT '24, September 04–06, 2024, Bremen, Germany

ACM ISBN 979-8-4007-1094-0/24/09
https://doi.org/10.1145/3677525.3678645

1 INTRODUCTION

Research has shown that artificial intelligence (AI) and machine learning (ML) systems, when trained without adequate consideration, often produce biased outcomes. This can lead to disparities in critical areas like college admissions, financial access, legal processes, and healthcare, potentially impacting human lives in detrimental ways [24]. Despite growing research on techniques to combat bias in AI, these methods have not yet gained significant traction in the actual deployment of AI systems across industries, government agencies, and the public sector. One reason for this is the complex nature of fairness. Achieving fairness in AI is a multi-faceted task, involving conflicting metrics and underlying societal assumptions that are challenging to balance. For instance, determining the extent to which racist and sexist societal processes cause harm and the appropriate application of affirmative corrections is not straightforward. Moreover, fairness metrics often contradict each other, creating a dilemma for developers and organizations [6]. There is also concern that focusing on fairness might compromise the accuracy of AI systems, which can discourage organizations from investing in fair AI systems since it may impact profitability.

A common first step in a fairness intervention is selecting fairness metrics, which reflect different social values and assumptions. Generally, AI fairness definitions can be categorized based on two main aspects: the granularity of protection and the treatment of disparities. *Individual fairness* ensures that comparable individuals receive similar treatment from the algorithm, while *group fairness* aims to ensure that various demographic groups (such as those based on race and gender) are treated similarly. *Infra-marginal* definitions accept existing disparities between demographic groups as legitimate, whereas *intersectional* definitions consider these disparities as unfair and requiring mitigation. In this work, we contend that *different stakeholders may prefer different fairness metrics*, and *optimizing a compromise between multiple fairness metrics and accuracy within a single system may afford a consensus solution.* To help stakeholders to address the challenges in choosing between different fairness metrics and balancing accuracy and fairness, we propose a *unifying human-centered fairness framework* that simplifies decision-making for the stakeholders. By formulating each metric within a single coherent and consistent framework, our approach reduces the learning curve for non-expert stakeholders. Our main contributions include:

(1) A unified fairness framework which systematically encodes multiple fairness metrics encompassing a broad spectrum of values and priorities in a consistent and elegant manner.
(2) A methodology for assigning relative weights to multiple fairness criteria and predictive accuracy, along with a learning algorithm to achieve the desired balance, and
(3) Experimental validation on two benchmark datasets.

2 BACKGROUND

Here, we discuss theories on fairness and their AI implementations. Our framework draws on two major conceptions of fairness which differ in their philosophical underpinnings and the treatment of disparities, *intersectionality* [9, 11, 28] from the humanities and legal literature, and *infra-marginality*, advanced by public policy, and law scholars [2, 27]. The concept of *intersectionality* was introduced by Kimberlé Crenshaw in 1980's [11] and later popularized by Patricia Hill Collins [9] in a broader context. Crenshaw advanced intersectionality as a lens to examine societal injustices, stemming from the perception that sexism and racism often overlap, resulting in greater harm to Black women than either issue alone would cause. The generalized extension advocated by Collins and others [9, 10] focuses on the systems of oppression built into society that lead to systematic disadvantages along intersecting dimensions such as race, sexual orientation, disability, and social class. It emphasizes how individuals can experience multiple forms of oppression simultaneously, resulting in complex and intersecting disadvantages, thus taking a politically progressive stance. Infra-marginality-based [27] notion of fairness, in contrast, operates under the premise that demographic groups inherently possess different distributions of risk or ability. It suggests that disparities in merit or risk among different groups should be considered valid, given the assumption that society provides a fair and level playing field. It assumes that disparities arise due to factors such as personal effort, talent, or choices rather than systemic societal inequalities. They contend that efforts to ensure equal outcomes overlook the inherent diversity among individuals and groups, potentially leading to misallocation of resources or unfair advantages to certain groups. However, critics of infra-marginality challenge this perspective. They argue that differences in merit or risk are often influenced by systemic structural disadvantages embedded in society. Factors such as racism, sexism, inter-generational poverty, and mass incarceration create barriers that limit opportunities for certain groups, leading to disparities. Individuals from marginalized communities may face challenges such as limited access to quality education, employment discrimination, or disproportionate policing, impacting their performance and outcomes [10–12, 16, 29].

Although the discussion about fairness and the treatment of disparity has so far focused on a group level, philosophers also view this idea on a more granular level of individuality. Calsamiglia [8] explains, "Philosophers define equality of opportunity as the requirement that an *individual's* well being be independent of his or her irrelevant characteristics."

Previous studies in AI fairness have laid the groundwork for understanding the associated complexities and challenges [3, 7, 13, 15, 23, 32]. One significant line of research has concentrated on defining and analyzing different fairness metrics. These metrics can broadly be categorized as either *individual* fairness metrics, which aim to ensure similar treatment by AI systems for similar individuals, and *group* fairness metrics, such as demographic parity, which aim to ensure equitable treatment of demographic groups [13]. Subgroup fairness metrics, which consider the intersections of demographic groups, have also been proposed [17, 19]. Other approaches include causal fairness [22], and fair representation learning [31]. Furthermore, researchers have explored the implications of fairness interventions on predictive performance [18, 21]. There is a growing understanding that promoting fairness often involves trade-offs with accuracy. Understanding these trade-offs is crucial for stakeholders when designing and deploying fair AI systems.

However, despite these advancements, a comprehensive framework that integrates diverse fairness metrics and allows stakeholders to assign weights to them is still lacking. While specific fairness metrics provide valuable insights into specific aspects of fairness, they often do not capture the full complexity of real-world fairness considerations. Although Berk et al. [5] did propose a *hybrid* fairness penalty term that incorporates individual and group notions of fairness, they didn't address the treatment of disparities (infra-marginality or intersectionality). Moreover, they did not take a human-centered approach, e.g., it did not provide stakeholders with the flexibility to choose weights according to their priorities, or help them achieve a consensus solution.

3 UNIFYING FAIRNESS FRAMEWORK

Achieving fairness in AI systems necessitates a comprehensive approach that considers multiple fairness metrics and the diverse values of stakeholders. Our proposed unifying fairness framework aims to address this complexity by allowing stakeholders to assign weights to their preferred fairness metrics. Multiple stakeholders cam compromise on the fairness weights, counteracting developers' "subjective bias" [25].

Our framework is partially inspired by the Apple Card debacle, where Steve Wozniak, Apple's co-founder, disclosed that the card granted him a credit limit ten times higher than that of his wife, despite their shared assets and similar credit scores.[1] Here, the alleged unfairness involved pairs of similar individuals from different demographics and we incorporated this concept into our proposed general framework.

3.1 Fairness Metrics

We propose a unifying fairness framework consisting of a collection of fairness metrics representing a spectrum of fairness notions, each formulated in a consistent manner using ratio-based formulations inspired by the 80%- rule or more generally, the p%-rule, a legal criterion for fairness [14]. Zafar et al. [30] investigated the application of p%-rule constraints using logistic regression and SVM algorithms. However, their method exhibited lower performance compared to our neural network-based classification algorithm. The use of ratio-based fairness in the legal system suggests that non-expert stakeholders can grasp these concepts. Our approach considers both infra-marginality and intersectionality perspectives, and recognizes the importance of considering fairness at both the individual and group levels, resulting in four fairness metrics (see Table 1). All four metrics are based on the ratios of a classifier's class probabilities, comparing unprivileged (y_j) to privileged (y_i) individuals or groups. The metrics differ in their granularity and in their treatment of disparities. *Infra-marginal individual fairness* takes the average ratio of the model's class probabilities for matched pairs of individuals (e.g., Steve Wozniak and his wife) based on a

[1] www.cnn.com/2019/11/10/business/goldman-sachs-apple-card-discrimination/

	Infra-marginal	Intersectional
Individual	$avg(Pr_M(y_i)/Pr_M(y_j))$, matched pairs $r_i^{(f)} \approx r_j^{(f)}$	$avg(Pr_M(y_i)/Pr_M(y_j))$ matched pairs with parity-adjusted $r^{(f)}$
Group	$avg(Pr_M(y_i))/avg(Pr_M(y_j))$ matched pairs $r_i^{(f)} \approx r_j^{(f)}$	$avg(Pr_M(y_i))/avg(Pr_M(y_j))$ unmatched instances

Table 1: The proposed framework systematically encodes the main types of fairness, categorized according to level of granularity and desired value systems. Here, $Pr_M(y)$ indicates classifier probabilities, i, j are from different demographics (e.g. male, female), and $avg(\cdot)$ computes the mean.

"fair risk score" $r^{(f)}$ which estimates class probabilities in a relatively fair manner. This "fair risk score" measures merit or risk based on the subset of the features that are deemed relatively fair (e.g., assets and credit scores), as we explain below. This metric reflects a politically conservative notion of fairness, as it does not account for the unfair societal processes that influence the scores, but rather assumes that the social context is a meritocratic "level playing-field." Conversely, *infra-marginal group fairness*, which is also considered conservative, calculates the ratio of the averages of probabilities for the matched pairs. For *intersectional group fairness*, the instances are unmatched, so no emphasis is given to their merit-based scores. Instead, their classifier's probabilities are averaged for both demographics, resulting in a ratio that is more inclusive of their differences in qualifications. In the case of *intersectional individual fairness*, this disparity is addressed by affirmatively adjusting the fair risk scores by adding a constant to all the unprivileged group's scores $r_j^{(f)}$ so that $avg(r_j^{(f)}) = avg(r_i^{(f)})$, the mean of the privileged group's scores, before matching unprivileged and privileged individuals. Thus, the *intersectional* notions of fairness address the unfair societal processes that influence the qualifications of individuals in unprivileged demographics.

3.2 Details of the Proposed Methodology

Our unifying framework focuses on comparison of *matched pairs of individuals*, based on a matching technique adapted from methods used in causal analysis such as *propensity score matching* [26]. We select a small number of baseline features (there could be just one), considered to be relatively fair (e.g. assets and credit scores in the Apple card example) and extracted for each individual n, $\mathbf{x}_n^{(f)}$. We train a logistic regression model on the baseline features to estimate "fair" risk scores for class $y = 1$ for each individual, $r_n^{(f)} \triangleq Pr(y = 1|\mathbf{x}_n^{(f)})$. These risk scores should be relatively "fair," although we acknowledge that they might be unreliable since not all the features in $\mathbf{x}$ were used. Then, individuals i from one demographic (e.g. men) are matched with an individual j from a different demographic (e.g. women) who are "similarly qualified" according to $r^{(f)}$. We then use the matched (and unmatched) pairs to construct the four ratio-based fairness metrics described in Table 1. In our general framework, each of these fairness metrics are given a *human-assigned weight* w_m, and the weighted metrics are added to construct an overall fairness metric $R(\mathbf{X};\theta)$ in Equation 1:

$$R(\mathbf{X};\theta) = \sum_{m=1}^{4} w_m R_m(\mathbf{X};\theta) \ . \tag{1}$$

Throughout the paper $w_{\text{I-M,ind}}$, $w_{\text{int,ind}}$, $w_{\text{I-M,grp}}$ and $w_{\text{int,grp}}$ will be used to represent the weights of the four metrics (I-M = infra-marginal, ind = individual, int = intersectional, and grp = group), and similarly for the fairness metrics, e.g., $R_{\text{I-M,ind}}(\mathbf{X};\theta)$. The overall combined fairness metric is weighted by the hyper-parameter λ based on the accuracy-fairness trade-off preference of the stakeholder. Details about the selection of λ and weights, w_m can be found in the results and case study section. Finally a learning algorithm enforces our fairness framework based on the following objective function:

$$\min_{\theta} f(\mathbf{X};\theta) \triangleq \frac{1}{N}\sum_{n=1}^{N} L(\mathbf{x}_n;\theta) - \lambda R(\mathbf{X};\theta) \ . \tag{2}$$

Here, N is the number of data points, $x_n \in \mathbf{X}$ is a data point in the training set, L is a loss function, R is a fairness metric (higher is better), θ is the model's parameters, and λ is a hyper-parameter that trades between the prediction loss and fairness.

4 EXPERIMENTAL SETUP

We evalauted our framework using the UCI Adult census dataset [4, 20] and the COMPAS recidivism dataset [1] (cf. the Appendix). These datasets are widely recognized as benchmarks for fair machine learning classification tasks. To calculate the "fair risk score" for each individual, we identified a set of baseline features that are considered relatively fair. For the UCI Adult dataset, we used *education level*, and for the COMPAS dataset, we used *prior offenses count*. We then trained a logistic regression model with these features.

5 RESULTS

Accuracy-Fairness Trade-Off: We analyzed the trade-offs between accuracy and different fairness metrics in our framework. In Figure 1, the x-axis represents fairness, while the y-axis shows the predictive accuracy of the classifier. Each dot represents the trade-off for different λ = [0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1]. As λ increases, indicating a higher importance given to fairness, the accuracy of the classifier decreases, while different metrics have different accuracy trade-offs. For example, in the UCI Adult dataset we see a significant drop in accuracy when the two individual fairness metrics are prioritized (λ>0.5). In the COMPAS dataset (shown in the Appendix), the trend is similar but varies in magnitude. This highlights how application contexts can influence the trade-off. For COMPAS, prioritizing fairness could mean the difference between unjust incarceration and fair treatment, while compromising too much on accuracy might risk harming public safety.

Fairness Metric Trade-Off: The plot in Figure 2 illustrates the trade-off between two fairness metrics: infra-marginal individual fairness (y-axis) and intersectional group fairness (x-axis) (a similar plot for the UCI Adult dataset is shown in the Appendix). As the weight shifts from infra-marginal individual fairness to intersectional group fairness, the trade-off curve indicates that improving one metric often compromises the other. This type of fairness-fairness plot applies generally but it addresses only two kinds of fairness metrics. In our framework, we have four fairness metrics (cf. Table 1), and each of these can be compared against the others. Each data point in the plot represents the model's fairness evaluation, specifically evaluated for $R_{I-M,ind}(\mathbf{X};\theta)$ and $R_{int,grp}(\mathbf{X};\theta)$, trained

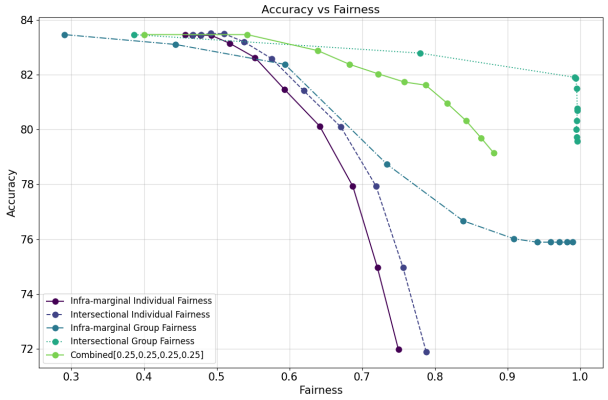

Figure 1: Accuracy-fairness trade-off for the fairness metrics (UCI Adult dataset). Each data-point represents a different λ.

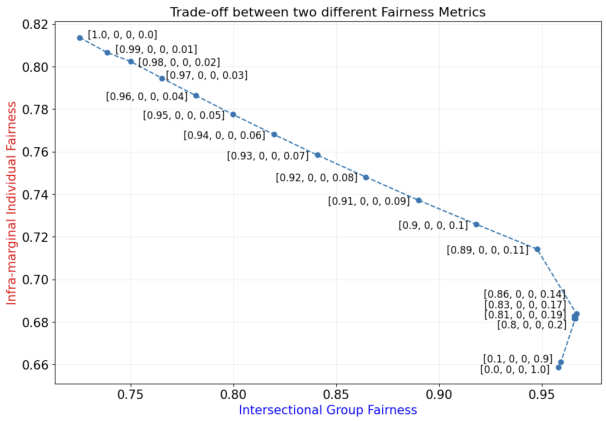

Figure 2: Trade-off between infra-marginal individual and intersectional group fairness (ProPublica COMPAS dataset).

on a specific weight configuration [$w_{\text{I-M,ind}}$, $w_{\text{int,ind}}$, $w_{\text{I-M,grp}}$, $w_{\text{int,grp}}$], with λ = 1.By adjusting these weights, stakeholders can visualize and understand the trade-offs between different fairness considerations, allowing for informed decision-making.

6 CASE STUDY: EXPLORING FAIRNESS IMPLICATIONS IN THE COMPAS DATASET

To illustrate the practical implications of our proposed fairness framework, we conducted a case study using the COMPAS dataset. We examined the application of fairness trade-offs in the bailing/sentencing of defendants based on their recidivism scores. In the Appendix, we include a second scenario: the allocation of social work labor resources for defendants. Both scenarios utilize outcomes derived from a fair machine learning model to inform decisions, guided by our fairness trade-off framework. In both scenarios, stakeholders could use trade-off plots similar to Fig. 2 to find an acceptable balance between their preferred fairness metrics.

Bailing/Sentencing of Defendants: In the context of bailing or sentencing a defendant, primary stakeholder groups include public safety advocates and civil rights advocates. Both groups rely on recidivism scores to assess the likelihood of re-offending, but they prioritize different aspects of fairness.

- *Public Safety Advocates* aim to prevent recidivism and reduce crime, prioritizing public safety, hence aligning more with infra-marginal individual fairness. This metric *matches* individuals based on qualifications or features, not accounting for societal biases that may influence these features. They might assign weights such as $w_{\text{I-M,ind}}$ = 0.95 and $w_{\text{int,grp}}$ = 0.05, with λ = 1.0. For this setting, accuracy of the model was 96.34%, $R_{I-M,ind}(\mathbf{X};\theta)$ was 0.78, $R_{int,grp}(\mathbf{X};\theta)$ was 0.80, and the overall weighted fairness metric $R(\mathbf{X};\theta)$ was 0.78.
- *Civil Rights Advocates* focus on preventing systemic injustices and ensuring fair treatment across demographic groups. They prioritize intersectional group fairness, which averages classifier probabilities across demographic groups, without *matching*, thus addressing broader societal inequities. They might assign higher weights to intersectional group fairness, such as $w_{\text{I-M,ind}}$ = 0.1 and $w_{\text{int,grp}}$ = 0.9. In this configuration, the accuracy dropped slightly to 96.18%, $R_{I-M,ind}(\mathbf{X};\theta)$ was 0.66, $R_{int,grp}(\mathbf{X};\theta)$ was 0.96, and the overall weighted fairness metric $R(\mathbf{X};\theta)$ was 0.93.

Assigning more weight to intersectional group fairness causes a minimal accuracy drop (≈0.16%) while significantly improving fairness from 0.80 to 0.96, compared to no fairness intervention where $R_{int,grp}(\mathbf{X};\theta)$ was 0.73. But it comes at the cost of infra-marginal individual fairness, dropping from 0.78 to 0.66. Therefore, a consensus solution may balance infra-marginal individual fairness and intersectional group fairness by setting the weights to, e.g., $w_{\text{I-M,ind}}$ = 0.92 and $w_{\text{int,grp}}$ = 0.08, with λ = 1.0, maintaining high accuracy (96.26%) while $R_{I-M,ind}(\mathbf{X};\theta)$ = 0.75, $R_{int,grp}(\mathbf{X};\theta)$ = 0.86, and overall weighted fairness metric $R(\mathbf{X};\theta)$ = 0.76. This configuration demonstrates an improvement in the two chosen fairness metrics over the typical model with no fairness intervention, where $R_{I-M,ind}(\mathbf{X};\theta)$ = 0.62, $R_{int,ind}(\mathbf{X};\theta)$ = 0.73, $R(\mathbf{X};\theta)$ = 0.68 and accuracy = 96.99%. This balance shows a willingness to sacrifice a small amount of **infra-marginal** individual fairness for a considerable gain in **intersectional** group fairness, a desirable trade-off in applications where mitigating systemic biases is critical.

7 CONCLUSION

We have proposed a human-centered fairness framework that allows stakeholders to navigate fairness complexities in AI systems by prioritizing fairness based on their preferences while balancing predictive accuracy. Experimental validation showed that adjusting fairness weights can control this balance and the importance of different fairness considerations. By refining and expanding our framework, we aim to support the development of fair and equitable AI systems for the benefit of all.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No.'s IIS1927486; IIS2046381. 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.

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica, May* 23 (2016).

[2] Ian Ayres. 2002. Outcome tests of racial disparities in police practices. *Justice research and Policy* 4, 1-2 (2002), 131–142.

[3] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.

[4] Barry Becker and Ronny Kohavi. 1996. Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

[5] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2017. A convex framework for fair regression. *4th Annual Workshop on Fairness, Accountability, and Transparency in Machine Learning. ArXiv preprint arXiv:1706.02409 [cs.LG]* (2017).

[6] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2021. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research* 50, 1 (2021), 3–44.

[7] Reuben Binns. 2018. Fairness in machine learning: Lessons from political philosophy. In *Conference on fairness, accountability and transparency*. PMLR, 149–159.

[8] Caterina Calsamiglia. 2005. *Decentralizing equality of opportunity and issues concerning the equality of educational opportunity*. Yale University.

[9] Patricia Hill Collins. 2022. *Black feminist thought: Knowledge, consciousness, and the politics of empowerment*. routledge.

[10] Combahee River Collective. 1978. A Black Feminist Statement. In *Capitalist Patriarchy and the Case for Socialist Feminism*, Zillah Eisenstein (Ed.). Monthly Review Press, New York.

[11] Kimberlé Crenshaw. 2013. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. In *Feminist legal theories*. Routledge, 23–51.

[12] Angela Y Davis. 2011. *Are prisons obsolete?* Seven stories press.

[13] 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*. 214–226.

[14] Equal Employment Opportunity Commission. 1978. Guidelines on Employee Selection Procedures. *C.F.R.* 29.1607 (1978).

[15] Moritz Hardt, Eric Price, Nati Srebro, et al. 2016. Equality of opportunity in supervised learning. In *Advances in NIPS*. 3315–3323.

[16] Bell Hooks. 2014. Ain't I a woman: Black women and feminism. (2014).

[17] Rashidul Islam, Kamrun Naher Keya, Shimei Pan, Anand D Sarwate, and James R Foulds. 2023. Differential fairness: an intersectional framework for fair AI. *Entropy* 25, 4 (2023), 660.

[18] Rashidul Islam, Shimei Pan, and James R. Foulds. 2021. Can We Obtain Fairness For Free?. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (Virtual Event, USA) *(AIES '21)*. Association for Computing Machinery, New York, NY, USA, 586–596. https://doi.org/10.1145/3461702.3462614

[19] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In *Proceedings of the 35th International Conference on Machine Learning, PMLR 80 (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, Stockholmsmässan, Stockholm Sweden, 2569–2577.

[20] Markelle Kelly, Rachel Longjohn, and Kolby Nottingham. 2023. The UCI machine learning repository. *URL https://archive.ics.uci.edu* (2023).

[21] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).

[22] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in NIPS*.

[23] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)* 54, 6 (2021), 1–35.

[24] Cathy O'neil. 2017. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

[25] Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. In *Proceedings of the aaai conference on artificial intelligence*, Vol. 34. 480–489.

[26] Paul R Rosenbaum and Donald B Rubin. 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* 39, 1 (1985), 33–38.

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

[28] Sojourner Truth. 1851. Ain't I A Woman? Speech delivered at Women's Rights Convention, Akron, Ohio.

[29] Johanna Wald and Daniel J Losen. 2003. Defining and redirecting a school-to-prison pipeline. *New directions for youth development* 2003, 99 (2003), 9–15.

[30] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. 2017. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*. 962–970.

[31] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International conference on machine learning*. PMLR, 325–333.

[32] Indrė Žliobaitė. 2017. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery* 31, 4 (2017), 1060–1089.