Are “Intersectionally Fair” AI Algorithms Really Fair to Women of Color? A Philosophical Analysis

Youjin Kong
Philosophy, Oregon State University, Corvallis, OR, USA
youjin.kong@oregonstate.edu

ABSTRACT

A growing number of studies on fairness in artificial intelligence (AI) use the notion of intersectionality to measure AI fairness. Most of these studies take intersectional fairness to be a matter of statistical parity among intersectional subgroups: an AI algorithm is “intersectionally fair” if the probability of the outcome is roughly the same across all subgroups defined by different combinations of the protected attributes. This paper identifies and examines three fundamental problems with this dominant interpretation of intersectional fairness in AI. First, the dominant approach is so preoccupied with the intersection of attributes/categories (e.g., race, gender) that it fails to address the intersection of oppression (e.g., racism, sexism), which is more central to intersectionality as a critical framework. Second, the dominant approach faces a dilemma between infinite regress and fairness gerrymandering: it either keeps splitting groups into smaller subgroups or arbitrarily selects protected groups. Lastly, the dominant view fails to capture what it really means for AI algorithms to be fair, in terms of both distributive and non-distributive fairness. I distinguish a strong sense of AI fairness from a weak sense that is prevalent in the literature, and conclude by envisioning paths towards strong intersectional fairness in AI.

CCS CONCEPTS

• **Computing methodologies** → Artificial intelligence; Philosophical/theoretical foundations of artificial intelligence; Machine learning; • **Social and professional topics** → User characteristics; Gender; User characteristics; Race and ethnicity.

KEYWORDS

Fairness and Bias in AI, Intersectionality, Philosophical Analysis of Fairness, Feminist and Critical Race Social Philosophy

ACM Reference Format:
Youjin Kong. 2022. Are “Intersectionally Fair” AI Algorithms Really Fair to Women of Color? A Philosophical Analysis. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3531146.3533114

FAccT ’22, June 21–24, 2022, Seoul, Republic of Korea
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9352-2/22/06…$15.00
https://doi.org/10.1145/3531146.3533114

1 INTRODUCTION: GROUP FAIRNESS, FAIRNESS GERRYMANDERING, AND INTERSECTIONAL FAIRNESS

In recent years, there has been an increasing concern about bias in artificial intelligence (AI). Studies have shown that AI algorithms, which are supposedly “neutral” and “objective” (or at least less biased than humans), actually reflect and reproduce racism, sexism, classism, and other forms of social injustice [4, 21, 42, 44]. In particular, a 2016 ProPublica article [1] revealed that COMPAS, a recidivism prediction algorithm widely used in US courtrooms, was biased against Black people.[1] In predicting who were likely to commit new crimes, the COMPAS algorithm tended to mark Black defendants as higher risk, falsely labeling them as future criminals twice more often than their white counterparts. In contrast, whites were labeled as lower risk but reoffended at twice the rate as Blacks. The racial disparities in risk scores are based on racial disparities in other sectors of the US society, such as the disproportionate policing and mass incarceration of Black people. As the racially biased algorithm is used in courts under the disguise of being “data-driven” and “impartial,” it reinforces racial injustice against Black people.

In response, AI researchers have taken great efforts to de-bias algorithms and improve fairness in AI.[2] The AI fairness literature has advanced multiple definitions of fairness that are often categorized as *group fairness.*[3] According to the group fairness notion (also

[1] I capitalize “Black” throughout this article, following the reasoning of critical race philosophers like Kwame Anthony Appiah [2]: “A good reason to capitalize the racial designation “black” ... is precisely that black, in this sense, is not a natural category but a social one—a collective identity—with a particular history. ... Giving *black* a big *B* could signal that it’s not a generic term for some feature of humanity but a name for a particular human-made entity.” To stress that races are “products of social forces,” Appiah also capitalizes “White.” However, the present article will not capitalize “white,” partly because it is common for white supremacist websites to capitalize White as a way to ennoble them, and mostly because “white people in general have much less shared history and culture, and don’t have the experience of being discriminated against because of skin color.” [3] The latter is the rationale for many news organizations’ decision to capitalize Black but not white, which they made in the wake of the police killing of George Floyd and Black Lives Matter protests. I will also capitalize Black but not white, in order to accentuate the history of racial discrimination against Black people that racially privileged (i.e., white) people have not had to go through. It is notable, though, some anti-racist scholars like Eve Ewing [22] have chosen to capitalize White to challenge the invisibility that enables white people to “get to be only normal, neutral, or without any race at all, while the rest of us are saddled with this unpleasant business of being racialized.” “When we ignore the specificity and significance of Whiteness,” Ewing continues, “we contribute to its seeming neutrality and thereby grant it power to maintain its invisibility.”

[2] Note that there is a tendency in AI fairness research to identify the notion of AI algorithms being “fair” with being “de-biased.” I will use the terms “unfairness” in AI and “bias” in AI interchangeably. However, as I will argue later in the paper, “fairness” in AI requires more than “de-biasing” algorithms.

[3] Another influential category is individual fairness, which stemmed from objections to the group fairness notion. Individual fairness says that “less qualified individuals should not be favored over more qualified individuals” for the purpose of achieving fairness between groups [5, pp. 515-516]. This paper will focus on group fairness as it is the notion of fairness directly related to the topic of this paper, namely, intersectional fairness in AI. For an overview of definitions of fairness in the AI literature, see [36, 41].

known as statistical fairness), fairness is defined as the equality of a statistical measure between "protected" (marginalized) and "unprotected" (privileged) groups. For example, a recidivism prediction algorithm is "fair" with respect to race, if the probability for Black defendants to be classified as future criminals is approximately the same as that for white defendants. This fairness standard, which seeks equality of the likelihood of positive outcomes, is called "statistical parity" or "demographic parity." Another group fairness standard seeks parity of error rates (such as false positive rates) between groups. If the rate of Black defendants to be mistaken as future criminals is the same as that of white defendants, the algorithm is "fair." In both standards, the key idea of group fairness is to "treat different groups equally." It appeals to the moral intuition that there should be no discrimination against certain groups based on their race, gender, class, and other identity categories [5, 32, 36].

One problem of the group fairness notion is that it is "only suited to a limited number of coarse-grained, prescribed protected groups" [5, p. 515]. Critics note that group fairness measures take only a single attribute (e.g., race) into account, and thus fail to consider subgroups defined by an intersecting combination of attributes (e.g., race and gender). Buolamwini and Gebru's groundbreaking work [6] drew attention to this problem of *intersectional bias* in AI. They found that major face recognition algorithms provided by Microsoft, IBM, and Face++ performed better on recognizing men than recognizing women (gender discrimination), better on people with lighter skin tones than people with darker skin tones (racial discrimination), and yielded the worst accuracy on women with darker skin tones (intersection of gender and racial discrimination). The error rates for Black women ranged from 20.8 – 34.7%, which were much higher than those for white men (0.0 to 0.3%), white women (1.7 – 7.1%), and Black men (0.7 – 12.0%).

It is noteworthy that the algorithms performed much worse on Black women, even compared to white women and Black men. Although not as accurate as recognizing white men, the algorithms did an acceptable job recognizing white women and Black men, especially when the error rates for the respective groups were as low as 1.7% and 0.7%. Yet for Black women, no algorithm did a decent job: all algorithms failed to recognize about 3 out of 10 Black female faces.[4] As Kimberlé Crenshaw [14] famously noted, the particular form of discrimination that Black women experience at the intersection of racism and sexism is "greater than the sum" of racism experienced by Black men and sexism experienced by white women.

Kearns and colleagues [32] coined the term "fairness gerrymandering" to explain this kind of intersectional unfairness in AI. The term refers to cases where the algorithm meets the fairness standard on each individual group but is unfair on their intersectional subgroups. The following toy example illustrates fairness gerrymandering:

> "Imagine a setting with two binary features, corresponding to race (say black and white) and gender (say male and female) ... Consider a classifier that labels an example positive if and only if it corresponds to a black man, or a white woman. Then the classifier will appear to be equitable when one considers either protected attribute alone, in the sense that it labels both men and women as positive 50% of the time, and labels both black and white individuals as positive 50% of the time. But if one looks at any conjunction of the two attributes (such as black women), then it is apparent that the classifier maximally violates the statistical parity fairness constraint." [32, p. 1]

In sum, Kearns et al as well as Buolamwini and Gebru highlight the limitation of the single-axis approach adopted by group fairness measures. Insofar as AI fairness research analyzes bias only along the axis of race or that of gender, it cannot adequately address intersectional bias experienced by Black women and other women of color. Feminist data scientists and information studies scholars thus have called for intersectional approaches to AI fairness [9, 13, 16, 28]. Their efforts have contributed to an emerging interest in intersectional fairness within non-explicitly feminist research as well. Several studies suggested technical solutions to fairness gerrymandering [27, 32, 34, 59], and a growing number of articles directly employ the concept of intersectionality for analyzing unfairness and improving fairness in AI [23, 31, 40, 50, 56, 57].

Although the gradual move away from single-axis approaches towards intersectional approaches is a positive development, the current ways that AI fairness research uses the notion of intersectionality entails problems. This paper critically analyzes the dominant interpretation of intersectional fairness in AI, and thereby aims to contribute to the efforts to build fairer AI algorithms. To this end, the following section starts by clarifying what the dominant view of intersectional fairness is. In section 3, I examine problems with the dominant approach. I argue that it misinterprets what intersectionality is and what fairness is, and thus fails to reach its original goals of rectifying intersectional bias in AI and creating fairer algorithms. Section 4 concludes by outlining ways forward for intersectional AI fairness research.

2 DOMINANT INTERPRETATION OF INTERSECTIONAL FAIRNESS: PARITY AMONG SUBGROUPS

Most studies in computer science that use intersectionality as a framework for measuring fairness take "intersectional fairness" to be a matter of ensuring statistical equality across subgroups. In their oft-cited work, Kearns and colleagues [32] propose what they call "rich subgroup fairness" as a solution to intersectional unfairness in AI. They argue to extend group fairness measures to "exponentially (or infinitely) many subgroups." This proposal is based on the recognition that group fairness notions that protect only one attribute (e.g., race) fail to address the intersection of multiple attributes (e.g., race and gender). Kearns et al contend that statistical measures such as rates to be classified or misclassified as x should be equalized not only between, e.g., Blacks and whites, but across, e.g., Black women, Black men, white women, and white men.

[4] In this paper, I use the terms "female/male" and "women/men" interchangeably. This is by no means to say that one's sex assigned at birth and their gender identity always correspond. Instead, this is to highlight "the continuous and dynamic relationships between biology, behavior, and social structures." [17] In the discussion of AI fairness, it is especially hard to separate sex (biological) from gender (social) and vice versa, as illustrated by face recognition algorithms that detect one's biophysical characteristics and classify their gender. In this regard, the use of the terms "female/male" and "women/men" in this paper is in line with van Anders's [55] neologism "gender/sex," which encompasses both gender and sex, as well as interactions between gender and sex.

If there are *n* binary protected attributes in measuring fairness, the number of intersectional subgroups that need to be considered is 2^n. This way, group fairness measures are to be implemented over an exponentially large collection of subgroups.

Foulds and colleagues [23] advance a similar notion of intersectional fairness. They refer to the so-called "four-fifths rule" [54] used as a guideline to enforce Title VII of the Civil Rights Act of 1964. The rule states that there is legal evidence of adverse impact if the selection rate for an underrepresented race, sex, or ethnic group is less than four-fifths of the rate for the most represented group. That is, the rule demands that the probability of (for example) women getting hired by an employer is no less than 80% of the probability of men getting hired. Foulds et al propose to apply this rule to multiple intersectional categories: "regardless of the combination of protected attributes, the probabilities of the outcomes [from an intersectionally fair mechanism] will be similar ... 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." [23, p. 5] In like manner, researchers seek intersectional fairness in AI by requiring statistical measures of their choice to be met on overlapping subgroups of the protected group [27, 34].

In short, the dominant approach to intersectional fairness in the computer science literature can be put as follows:

Definition 1. ***(intersectional fairness - PA).*** *An AI algorithm is intersectionally fair if it achieves* ***parity*** *of a statistical measure (e.g., 1.1, 1.2) among intersectional subgroups that are defined by different combinations of the protected* ***attributes****.*

For example, a loan approval algorithm is intersectionally fair with respect to gender, race, and nationality:

1.1 (statistical parity). if the probability of getting a loan is roughly the same across all subgroups defined by cross-cutting categories (e.g., women/men × Black/white × non-US/US = 8 subgroups); or

1.2 (equal false negative rates). if the rate for, e.g., Black women with non-US citizenship who have the ability to repay their loan are falsely denied a loan is roughly the same as the rate for other intersectional groups, such as white men with non-US citizenship and Black women with the US citizenship.

I will call this dominant view of intersectional fairness "PA," which stands for its two keywords: "parity" and "attributes." Unless otherwise specified, I will discuss the first sub-type of PA that demands equal probabilities of outcomes across the board (1.1 statistical parity), as it is one of the most common fairness measures used in the literature.

3 THREE PROBLEMS WITH THE DOMINANT INTERPRETATION

In this section, I identify and analyze three problems with PA: (1) an overemphasis on intersections of protected attributes, (2) an infinite regress and a reinscription of fairness gerrymandering, and (3) a narrow understanding of fairness as equal distribution.

3.1 Splitting subgroups along the line of identity categories

First, PA is so preoccupied with the intersection of identity categories (e.g., race, gender, and disability) that it fails to address the intersection of oppression (e.g., racism, sexism, and ableism), which is more central to intersectionality as a critical framework. In PA, intersectionality is seen a matter of splitting a group into finer subgroups along the lines of identity categories, or as AI researchers call them, "protected attributes." This view of is clearly expressed by Foulds and Pan, when they contend that one can mitigate intersectional bias in AI by "simply defining more fine-grained protected groups, e.g., designating Black women as protected" [24, p. 65]. They note that Kearns et al [32] and Foulds et al [23], the works examined above as examples of PA, successfully improve fairness by "enforc[ing] parity for groups at the intersection of the protected attributes." [24, p. 65] In other words, the path to intersectional fairness starts with combining multiple categories to define smaller subgroups so that one can apply fairness measures to all of the subgroups.

This is, however, a narrow interpretation of intersectionality, if not a misinterpretation. Intersectionality pertains not merely to *identity categories* but also and more to *structural oppression*. Structural analyses have characterized intersectionality since its inception, as noted by Collins and Bilge [11]. The Combahee River Collective Statement in 1977 [12], one of the early works that prompted the development of intersectional feminisms, pointed out that the "major systems of oppression"—racial, sexual, heterosexual, and class oppression—were "interlocking" to create the conditions of the lives of Black women. The same goes for today's discussion on what intersectionality is. While there is no one universal definition of intersectionality, feminist philosophers and theorists widely agree that "intersectionality's raison d'être lies in its attentiveness to power relations and social inequalities." [10, p. 3] Cho, Crenshaw, and McCall elaborate on the centrality of power in intersectionality as follows:

> "[W]hat makes an analysis intersectional is not its use of the term "intersectionality" ... [but] its adoption of an intersectional way of thinking about the problem of sameness and difference and its relation to power. This framing—conceiving of categories not as distinct but as always permeated by other categories, fluid and changing, always in the process of creating and being created by dynamics of power—emphasizes what intersectionality does rather than what intersectionality is. ... [I]ntersectionality helps reveal *how power works* in diffuse and differentiated ways *through the creation and deployment of overlapping identity categories*." [8, pp. 795, 797, emphasis added]

This is not to say that AI fairness research should abandon the language of identity, category, or attribute altogether. This is to say that identity should be examined in its relationship to power, rather than as an independent, self-sufficient unit of analysis in a vacuum. Black women are oppressed not because they have intersecting identities of "Black" and "women" per se, but because these identities are *shaped by and lived in the intersecting structure of racism and sexism.*

To illustrate, let us return to the case of face recognition algorithms that perform worst on Black women [6]. What causes this intersectional bias? One answer would be that datasets on which algorithms are trained have only few images of Black women. Existing datasets that serve as benchmarks for face recognition are composed mostly of whites and men [6, pp. 3, 6]. As the dataset has a large number of images of white men, the algorithm can easily learn to recognize their faces, which results in the best accuracy on this group. In contrast, the algorithm is not given many opportunities to learn how to classify Black women's faces accurately, which incurs the highest error rates on them.

Then what is the reason that datasets are so white- and male-dominated? Images for datasets are often collected by crowdworkers, and it is possible that these crowdworkers are biased. Like many other people living in the system of white supremacist patriarchy, they may have inadvertently identified "humans" with whites and/or males, or at least, regarded them as the representative sample of humans. When these crowdworkers participate in the task of collecting images of "human faces," they could select more white faces than those of Blacks and other people of color, and more male faces than those of women and nonbinary people. Crowdworkers might not even notice that they have collected few faces of Black women because it was not their intention, and yet, Black women become underrepresented in the resulting dataset.

The unintentional bias is also prevalent among AI engineers. There is a lack of racial and gender diversity in Big Tech companies such as Google, Facebook, and Microsoft. Not only are Black women underemployed as developers, but, as Noble points out, "jobs that could employ the expertise of people who understand the ramifications of racist and sexist stereotyping and misrepresentation and that require undergraduate and advanced degrees in ethnic, Black / African American, women and gender, American Indian, or Asian American studies are nonexistent." [42, pp. 69-70] It would be difficult for a team constituted mostly of white male developers who have not learned about the histories of racial bias in the US to notice the "blind spots" in datasets [29]. The team would proceed with the skewed dataset and train their face recognition algorithm with it, resulting in the biased algorithm that fails to recognize Black women.

As this example shows, most engineers and crowdworkers do not deliberately discriminate against Black women. The problem is rather that they are simply doing their jobs but their actions contribute to reproducing oppression. In the case of crowdworkers in particular, it is important to note that these workers are at the bottom end of the technical labor market. Based on their interviews with crowdworkers in Argentina and Venezuela, Miceli and colleagues emphasize that issues of so-called "worker bias" are in fact "manifestations of broader power asymmetries that fundamentally shape data: power asymmetries that are as trivial as being the boss in a tech company and have decision-making power, or being an underpaid crowdworker who risks being banned from the platform if they do not follow instructions." [37, p. 6]

In sum, the implicit bias of people who participate in the development process, the lack of workplace diversity, and the hierarchal and colonial labor market all operate together to result in biased algorithms. It is in this sense that Iris Marion Young conceptualized oppression as a "structure" or "system." Structural oppression, according to Young, is:

> "embedded in unquestioned norms, habits, and symbols, in the assumptions underlying institutional rules and the collective consequences of following those rules. ... In this extended structural sense oppression refers to the vast and deep injustices some groups suffer as a consequence of often unconscious assumptions and reactions of well-meaning people in ordinary interactions, media and cultural stereotypes, and structural features of bureaucratic hierarchies and market mechanisms—in short, the normal processes of everyday life." [58, p. 41]

At the core of intersectionality is the idea that multiple forms of oppression intersect to constitute such a *structure*. This is clearly illustrated by Kimberlé Crenshaw's intersectional analysis of domestic violence against women of color [14, 15]. Crenshaw, the Black feminist legal scholar who coined the term intersectionality, notes that intervention strategies based the experiences of white middle-class women do not help to address the unique challenges facing women of color. A better intervention should investigate how poverty, child care, lack of job skills, racially biased employment practices, and lack of language access and translated resources are intertwined and make it hard for many women of color to leave abusive relationships. In other words, efforts to create intersectionally fair strategies should start from examining how structural oppressions—i.e., class, gender, race, and language oppressions that are normalized as everyday life conditions of US society—intersect in women's experiences of domestic violence [15, pp. 1245-1251]. Without such examination, merely splitting women into finer groups along the lines of identity categories (e.g., poor, Asian, immigrant, less-English-proficient women) does not help.

Similarly, to build a face recognition algorithm that is intersectionally fair, the focus should be placed on the ways in which the intersectional structure of racism and sexism manifests throughout the crowdsourcing and AI development pipeline and is reinforced by biased algorithms. PA falsely suggests that intersectional bias in AI can be removed by simply generating more fine-grained combinations of identity categories or *attributes*. By doing so, PA diverts attention from what is more to the point: using intersectionality as a framework for analyzing and challenging complex systems of *oppression*.

3.2 Falling into either regress or gerrymandering

The predominant focus of PA on attributes leads to another problem: PA either faces an unwanted challenge of infinite regress, or it reinscribes the problem it claims to solve, namely, fairness gerrymandering.[5] A hypothetical case (modified from [18, 45]) would help illustrate the problem. In this case, a company uses an AI

[5] Both "regress" and "reinscription" are rhetoric to which critics of intersectionality (especially those in the humanities and social sciences) often appeal. PA adopts a narrow notion of intersectionality that focuses only on identity categories, which makes it prone to the regress critique and the reinscription critique. For detailed discussions of the two strands of critique, see [7, 52].

algorithm for its hiring process. The task of this algorithm is to predict which applicants are most likely to be successful when hired. (Here, success is defined as working at this company for at least four years and being promoted at least once.) To best predict it, the algorithm has been trained through the past 25 years of applications to the company. After a trial run, the company notices that the algorithm is biased in favor of white men and against Asian women. It is because white men have constituted the majority of the "successful" employees at the company. In contrast, Asian women employees were hardly ever promoted, which led many of them to leave the company. Trained with these data, the hiring algorithm has taught itself that white men are more likely to be successful employees and thus white male applicants are preferable to Asian female applicants.

How can this intersectional bias be corrected? PA's solution is to divide applicants along racial and gender lines and ensure that all racial-gender groups have an approximately equal chance of getting hired. Although this sounds straightforward, there is a myriad of factors that lead to a certain outcome from an AI algorithm. In the case at hand, not only the applicant's race and gender, but their age, disability, education level, first language, sexual orientation, and so on could intersect to affect the algorithm's prediction. In other words, all of these may be attributes that need protection. And there is an endless list of potential "protected attributes."[6] In addition to traditional identity categories, more concrete factors may have historically affected employee success at the company. Consider these two groups:

- AM1. Asian men who are work visa holders and graduates of US East Coast universities
- AM2. Asian men who are US citizens and graduates of US West Coast universities

Suppose that most employees from AM1 were "successful" at the company, whereas only a small number of AM2 employees met this standard of success. Then, in the same way as above, the algorithm would teach itself that AM1 applicants are preferable to AM2 applicants. This raises a number of questions. Should AM2 also be defined as a protected group like Asian women? Or, is it just a coincidence that employees from AM2 were less successful at the company, whereas the underrepresentation of Asian women employees is due to racial and gender discrimination, so AM2 does not need protection? To put it another way, is it sufficient to designate race and gender as protected attributes, or should racial-gender groups be split into even finer subgroups along the lines of visa status, location of alma mater, and so on? In responding to these questions of "how fine-grained protected subgroups should be," PA faces a dilemma between infinite regress and fairness gerrymandering.

On the one hand, if PA seeks parity among *all* possible subgroups, it would need to keep splitting groups into finer subgroups, until the point where there is no group and the individual is the only available unit of analysis. This is the problem of *infinite regress* that intersectionality has been frequently charged with [20]. Critics contend that intersectionality falls into an infinite regress in the following manner. According to intersectionality, claims about the nature of "black oppression" are misleading because:

> "what it means to be oppressed in virtue of blackness differs for black men and black women. By the same token, however, "black women's oppression" isn't a genuine kind either, because gender, race, and class intersect: what it means to be oppressed in virtue of black-womanhood differs for rich and poor black women. The same goes for sexuality, ability, religion, and a host of other significant social categories, potentially *ad infinitum*." [25, p. 1304]

While the infinite regress problem is often raised by philosopher critics, it is not a hypothetical problem that arises only on the theoretical level. Infinite regress poses actual challenges to engineers. If AI attempts to assess bias in all the exponentially many subgroups, it confronts at least two computational problems.

First, there would be too many subgroups to consider. When there are n binary protected attributes (e.g., Black/white, women/men, with/without disability), the number of subgroups to be taken into consideration is 2^n. If groups keep splitting along the lines of 30 attributes, the algorithm should assess bias and seek parity among 2^{30} = more than a billion intersectional subgroups. The number increases when protected features have more than two possible values (e.g., race: Black, White, Latinx, Native American/Indigenous, Asian, and so on). This way, the proposal to consider every possible subgroup becomes computationally impractical [5, 59].

Second, the size of subgroups would be too small. If subgroups are defined with combinations of 30 binary attributes, each of the 2^{30} groups may have only a couple of members, and sometimes no member at all. This raises a data scarcity issue [40, 56]. How does an algorithm seek parity among subgroups if there are too little data available for most of the groups? It also risks overfitting [32], the problem that an algorithm fits the given dataset so closely that it cannot be used for other datasets.

Taken together, the infinite regress problem suggests that PA is not a meaningful measure of fairness for marginalized (sub)groups. AI fairness researchers started employing the concept of intersectionality to mitigate bias against multiply-oppressed groups, such as women of color. However, the attempt to consider all subgroups defined by intersecting combinations of protected attributes leaves algorithms with too many and too small subgroups—or, more precisely, *individuals*. If what PA fundamentally seeks is statistical parity between individuals (e.g., all individual job applicants have an equal probability of getting hired), it is unclear how this race-blind/gender-blind fairness measure specifically tackles intersectional race and gender bias against women of color.

[6] I use quotation marks here to note that "protected attributes" are not natural categories. One suggestion I have received from commentators is to take only *legally* protected attributes into consideration. While this approach might help to prevent an infinite regress, it could exclude attributes that need protection but are not legally designated as protected attributes (yet). For example, the idea that one should not be discriminated against on the basis of sexual orientation and gender identity is a relatively new one, in both the history of the US anti-discrimination law and that of the international human rights law. Moreover, sexual orientation and gender identity are themselves fluid and contested, not fixed, attributes. My main point here is that researchers should be more critical of the notion of "protected attributes": what attributes are included in this notion and what are excluded. Instead of simply taking into account "legally protected" attributes and (as I will discuss below) "statistically relevant" ones, researchers must take note of how certain types of oppression remain marginalized from the discussion of the "protected attributes." In the present section, I will examine how "nonbinary" gender is excluded from sociotechnical systems, and demonstrate that *protection* and *relevance* are political issues.

To avoid an infinite regress, some researchers propose to take only "relevant" subgroups into consideration. Their line of argument proceeds as follows: In order to resolve both the problem of intersectional bias in AI (i.e., considering too few groups) and that of infinite regress (i.e., considering too many groups), we should split protected groups more finely but stop splitting at a "reasonable" point—somewhere between too few and too many. The question then is: What is a reasonable point and how do we know it? Kearns and colleagues make clear that when they ask for fairness across "exponentially many" subgroups, they do not refer to every possible subgroup. They refer only to "large structured" subgroups, for which "the dataset is sufficiently large" and "the statistical problem of learning and auditing fair classifiers is easy" [32, p. 2]. Hébert-Johnson and colleagues also require that algorithms be unbiased on all "efficiently- and computationally-identifiable" subgroups [27]. In short, according to these researchers, PA is to consider only *statistically meaningful* subgroups that can be identified by computers, as opposed to every possible subset.

This would lead to failure to protect minority subgroups [23]. Subgroups that are severely underrepresented in the dataset due to the very intersectional oppression they suffer are likely to be dismissed as "statistically meaningless." A more fundamental problem is why, after all, only "statistically meaningful" subgroups deserves protection. I argue that PA's presupposition that it is okay to exclude statistically less important subgroups from the fairness consideration is another form of fairness gerrymandering, namely, an *arbitrary* selection of protected attributes. As explained earlier, fairness gerrymandering refers to an algorithm being fair with respect to one attribute (e.g., statistical parity between Blacks and whites) but unfair with respect to multiple attributes (e.g., no statistical parity between Black women and other racial-gender subgroups). That is, fairness gerrymandering occurs "when we only look for unfairness over a small number of pre-defined groups" that are arbitrarily selected [32, p. 1]. PA seems to reinscribe this problem when it demands fairness measures to be met only across statistically meaningful subgroups. Just as algorithms that require fairness only among racial groups can be criticized for being arbitrary from the perspective of intersectional racial and gender injustice, algorithms that designate race and gender as protected attributes can also be criticized for being arbitrary from the perspective of intersectional race, gender, class, and disability injustice.

As such, any decision to delineate between relevant and irrelevant subgroups is susceptible to criticism that it is arbitrary. The response that "we protect only such-and-such attributes because they are statistically meaningful while others are not" is inadequate to resolve the arbitrariness/gerrymandering concern, since the question of what is "relevant" is itself a *political* battleground, not merely a *statistical* problem. Sasha Costanza-Chock's #TravelingWhileTrans anecdote illustrates this point clearly. Costanza-Chock discusses their lived experience as a white "gender nonconforming, nonbinary trans feminine person" in airport security lines:

> "[W]hen I [enter] the scanner, the TSA operator on the other side [is] prompted by the UI to select 'Male' or 'Female.' Since my gender presentation is nonbinary femme, usually the operator selects 'female.' However, the three dimensional contours of my body, at millimeter resolution, differ from the statistical norm of 'female bodies' as understood by the dataset and risk algorithm designed by the manufacturer of the millimeter wave scanner (and its subcontractors), and as trained by a small army of clickworkers tasked with labelling and classification ... If the agent selects 'male,' my breasts are large enough, statistically speaking, in comparison to the normative 'male' body-shape construct in the database, to trigger an anomalous warning and a highlight around my chest area ... In other words, I can't win. I'm sure to be marked as 'risky,' and that will trigger an escalation to the next level in the TSA security protocol." [13, part 1]

This experience shows how structural cisnormativity is built into scanning/risk detection algorithms and reproduced by marking trans and nonbinary people as the "risky Other." And cisnormativity is not the only type of oppression encoded in scanning technology. According to a report by ProPublica, scanners are prone to falsely identify Afros, braids, twists, and other hairstyles popular among Black women as signs of risk [35]. As cisnormativity, racism, and (hetero)sexism intersect, scanners would have an even higher false alarm rate when the passenger is a Black nonbinary person. Then what is the solution to this intersectional bias in the risk detection algorithm? PA would suggest enforcing parity of false alarm rates between intersectional subgroups defined by combinations of *gender* = {nonbinary, female, male} and *race* = {Black, white, other races}. The algorithm is "intersectionally fair" if the rate for Black nonbinary people to be falsely marked as risky is roughly the same as, for example, the rate for white females.

This fairness definition misses the point. To divide people into gender-racial subgroups and assess the false positive rate for Black nonbinary people, "nonbinary" must exist as one of the types of gender (and "Black" as a type of race). However, as Costanza-Chock's experience shows, the problem is that there is no such thing as "nonbinary gender" in the algorithm in the first place. *"Nonbinary" has not even been considered as a type of gender whose (statistical) relevance can be measured.* It is an "unobserved characteristic," in Tomasev et al's [51] terminology. Tomasev et al note that characteristics such as gender identity and sexual orientation are frequently unobserved, not merely due to their physical unobservability but also and more to the sociotechnical systems that exclude non-normative genders [51, pp. 260-261]. In order for "nonbinary" to be included and to count as a type of gender (as "female" and "male" do), there must be growing awareness in society—and among crowdworkers, engineers, and TSA agents—that gender is not a binary attribute, that cisnormativity operates as a structural oppression, and that it intersects with racial oppression in the lives of nonbinary people of color. Without such recognition, workers would keep labeling humans as either male or female, and the algorithm trained using such datasets would keep giving false alarms for people who do not fit within either label. These issues of relevance—or more precisely, what can be assessed as more or less relevant, and what remains unobserved from the consideration of relevance—are political problems, not purely statistical ones that computers can identify and measure.

I agree with Selbst and colleagues that the fairness literature often “fail[s] to account for the full meaning of social concepts such as fairness, which can be procedural, contextual, and contestable, and cannot be resolved through mathematical formalisms” [48, p. 61]. In the next section, I will elaborate on this idea of fairness being more than just mathematical formula.

3.3 Parity does not equal fairness

So far, this paper has focused on PA’s misinterpretation of intersectionality. The current section draws attention to PA’s misinterpretation of fairness. I argue that PA fails to capture what it really means for AI algorithms to be “fair,” in terms of both distributive justice and non-distributive justice.

Distributive justice is an umbrella term for theories that view justice as a proper distribution of benefits and burdens among members of society. Different branches hold different views of what is a “proper” distribution (see, for example, [19, 47]). PA takes a distributive approach to fairness. Suppose that a philosophy graduate program uses an AI algorithm to make admissions decisions. According to PA, this algorithm is intersectionally fair if it achieves parity of a statistical measure (e.g., the probability of being admitted) among intersectional subgroups of applicants (e.g., Black women, Black men, white women, and white men). That is, PA holds the view that *equal* distribution is a proper way of distributing benefits: the algorithm is fair if it distributes admissions rates equally across racial-gender groups.

However, there are cases in which *unequal* distribution is a more proper distribution. In order to undo the effects of structural oppression, I argue, AI-driven decision-making process may require more active interventions that allow a higher probability of preferable outcomes for marginalized subgroups than for privileged subgroups. This is especially the case when the marginalized are severely underrepresented. As is well known, philosophy in the US is a white- and male-dominated field, where Black women and other women of color are significantly underrepresented. According to a nationwide survey conducted in 2018 [30], Black people constituted only 1.1% of PhDs in philosophy, while making up 13.4% of the US populations and 8.8% of all doctorates. It was estimated in 2016 [26, 46] that only approximately 40 Black women had ever earned philosophy PhDs in the US. The dearth of Black women in philosophy is systemically reproduced through a number of mechanisms. There are few Black women philosophy professors who can serve as mentors and help Black women students navigate challenge during their graduate studies. Philosophical topics that Black women students may find interesting (e.g., Black feminist thoughts) would not be taken seriously in white male dominated philosophy departments, which might lead them to drop the program. The academic culture of US philosophy departments that is characterized by the lack of diversity and inclusivity may discourage Black women undergraduates from applying to graduate programs. In sum, there is a structural pattern in in the field of philosophy that marginalizes Black women.

Keeping this in view, let us return to the admissions algorithm example. It would be reasonable to assume that whites, especially white men, make up the majority of the applicants, while Black women constitute the minority. Suppose that there are 30 white male applicants and 3 Black female applicants to this graduate program. By PA’s definition, the algorithm is intersectionally fair if all racial-gender groups have equal chances (say, 33%) of getting in—that is, if 10 white men and 1 Black woman are likely to be accepted. Is it, though, really “fair”? The equalization of admissions rates does not ensure fairness, as it reflects and reproduces the status quo underrepresentation of Black women in philosophy. In order to actively mitigate the effects of white- and male-dominated culture of philosophy that has been hostile to Black women, the admissions algorithm may need to distribute a higher probability to Black women (for example, 66% = 2 out of 3 applicants get accepted, or even near 100% = 3 out of 3) than to white men. White male applicants have the systemic privilege of not having to go through the aforementioned challenges and disadvantages that Black female applicants experience. Without attending to the structural pattern that privileges certain groups while marginalizing others, a mere equal probability distribution does not achieve fairness.

This idea is in line with Iris Young’s critique of distributive justice. Analyzing problems of the distributive paradigm, Young argues that “the concepts of domination and oppression, rather than the concept of distribution, should be the starting point for a conception of social justice.” [58, p. 16] Young proposes a non-distributive paradigm that defines justice as the *elimination of structural oppression* [58, Ch. 1]. I argue that AI fairness should be examined through this lens of non-distributive justice. The notion of fairness as a matter of challenging and subverting oppression, as opposed to a matter of achieving statistical parity, better captures what it means for algorithms to be “fair” and provides better directions for developing “fairer” algorithms. In the remainder of this section, I discuss how PA is confronted with the two problems of the distributive paradigm that Young points out. In the next section I envision paths towards non-distributive intersectional fairness in AI.

First, Young maintains that the distributive paradigm obscures institutional contexts that shape the distributive pattern [58, pp. 18-23]. PA repeats this problem by taking a distributive approach to fairness. Consider, for example, loan approval algorithms. AI fairness researchers have worked towards developing algorithms that do not reproduce bias in current lending practices. Mortgage lending is notoriously discriminatory towards people of color and queer people. In 2017, Black applicants were rejected (18.4%) more than twice as often as white applicants (8.8%) [53]. A 2019 research article found that non-heterosexual couples were 73% more likely to be denied mortgages than otherwise similar heterosexual couples [49]. PA seeks to mitigate these biases by equally distributing some number (e.g., approval rates or false denial rates) across groups defined by combinations of race and sexuality. However, the mortgage denial gap is not just a matter of numbers; it has been caused by social, institutional, and historical contexts, including the stereotype of homeowners as white married heterosexual men, the credit score system working in favor of landlords, and socioeconomic inequality between races that stem from the history of slavery and Jim Crow [38, 39]. It is these *non-distributive* contexts that shapes the uneven, unfair *distributive* pattern of denial rates. Although seeking a more evenly distributed pattern is desirable, it is misleading to say that an algorithm’s fairness depends only on how it distributes the number. The exclusive focus on distribution “inappropriately restricts the scope of [fairness]” because, as Young points out, it fails to bring the non-distributive structures that determine the distributive pattern under scrutiny [58, p. 20].

Another problem with the distributive paradigm it overextends the concept of distribution to non-material values [58, pp. 24-30]. PA also understands non-material values, such as opportunities for education and employment, as if they were tangible material goods that can be allocated between groups. According to Young, this is a misconception of opportunity:

> "Opportunity is a concept of enablement rather than possession; it refers to doing more than having. A person has opportunities if he or she is not constrained from doing things, and lives under the enabling conditions for doing them. ... Evaluating social justice according to whether persons have opportunities, therefore, must involve evaluating not a distributive outcome but the social structures that enable or constrain the individuals in relevant situations." [58, p. 26]

Echoing this view, Hoffman [28] criticizes AI fairness research for examining discrimination only when it relates to particular distributive outcomes (in, for example, admission and hiring). In so doing, studies fail to address non-distributive, non-material type of discrimination. One such example is racist and sexist stereotypes built into search engines, which Noble's book *Algorithms of Oppression* [42] discusses in detail. Noble's Google search on the keyword "black girls" presented mostly pornographic websites, although the words "porn" or "sex" were not included in the query. The results that Google offered for the search on "unprofessional hairstyles for work" were images of Black women, while the search on "professional hairstyles for work" featured white women. Similarly, the "three black teenagers" search led to mug shots, whereas "three white teenagers" were represented as wholesome [42, Ch. 2].[7] As Hoffman aptly notes, these "representational and intimate harms are not easily or intuitively remedied. ... Money lost can be replaced and rights violated can be restored, but corporate apologies, subtle tweaks to a system, or even final compensation ring hollow in the face of attacks on one's dignity." [28, p. 908] This example again illustrates that fairness is more than a matter of distribution.

4 WAYS FORWARD: FROM WEAK TO STRONG FAIRNESS IN AI

In this paper, I have examined the dominant view in the AI fairness literature (PA) that an AI algorithm is intersectionally fair if it achieves statistical parity across intersectional subgroups. I have discussed three problems with this view. First, PA is so preoccupied with intersecting combinations of attributes such as race and gender that it diverts attention away from how racism, sexism, and other forms of oppression intersect to create bias in AI. Second, the preoccupation with attributes leads to a dilemma. PA either keeps dividing protected groups into ever finer subgroups along lines of multiple intersecting attributes, or it stops the regress at some point that can always be deemed an arbitrary selection. Lastly, by adopting a narrow understanding of fairness as equal distribution of outcomes, PA fails to address non-distributive aspects of fairness.

This paper as a whole is a response to Noble and Tynes's call for "intersectional critical race technology studies" (ICRTS). Noble and Tynes describe ICRTS as a research approach that:

> "interrogate[s] naturalized notions of the impartiality of hardware and software and what the Web means in differential ways that are imbued with power ... [and] examine[s] how information, records, and evidence can have greater consequences for those who are marginalized. Unequal and typically oppressive power relations map to offline social relations in ways that are often, if not mostly, predicated on racialized and gendered practices." [43, pp. 3-4]

I have taken this critical approach to a meta-analysis of the AI fairness literature that analyzes fairness in AI algorithms. Many studies in the literature start from the critical recognition that AI algorithms are not impartial and disproportionately affect the marginalized. In other words, AI fairness researchers sympathize with the call for ICRTS. Nevertheless, the actual practice of AI fairness research has been insufficient to realize this idea. The literature has focused too much on technological solutions to intersectional bias in AI and too little on how AI algorithms, as part of the intersectional structure of oppression, reflect and reinforce this structure.

I conclude by proposing future directions for intersectional AI fairness research. Rather than advancing a single methodology for all studies, I outline several questions that could help each study develop their own methods to move forward. To begin with, I suggest shifting the focus of fairness research from intersections of protected attributes to intersections of structural oppression. In the case of algorithms that are biased against Black women, such as the face recognition algorithm that performs worst on Black women, research for fairer algorithms can be guided by the following questions: Through what process is the structure of racial patriarchy is being *embedded* into AI algorithms? How does the biased algorithm *perpetuate* the racial patriarchy of society? In order to *resist* this intersecting structure of racial and gender oppression, how should the entire development process be redesigned? In this regard, I suggest that intersectional fairness research center more around non-distributive fairness than around distributive fairness. It is racist-sexist systems and contexts of society (i.e., non-distributive matters) that produces the unfair distributive pattern of error rates from the face recognition algorithm between Black women and white men. As such, the sole focus on equal distribution of error rates is a band-aid solution that keeps the underlying system intact.

Taking these points together, I invite AI fairness researchers to rethink what "fairness" is. Here I propose to distinguish between weak fairness and strong fairness. In a weak sense, AI fairness means passively and retroactively "de-biasing" the algorithm. This has been the focus of most studies in the literature, including those adopting PA. PA starts by detecting "blind spots": for example, an algorithm has been known to offer equal classification rates between women and men and between Blacks and whites, but it is found that there is no statistical parity between Black women and other racial-gender groups. Then it proceeds to "fix" the problem by applying the statistical parity standard not only to Blacks/whites and women/men, but to their intersectional subgroups. While debiasing using an equal distribution among intersectional subgroups

[7] After these search results went viral on social media, Google fixed them.

is a step forward to intersectional fairness in AI, I maintain that this alone cannot make AI algorithms substantively (as opposed to merely formally) fair. In that racial, gender, and other systemic oppression is being embedded into and reproduced by AI algorithms, making algorithms substantively fair involves resisting and undermining the very systems of oppression that create AI bias in the first place. Thus, AI fairness in a stronger sense means using *algorithms to actively and proactively challenge oppression and make society fairer*. A central guiding question for strong AI fairness is how to design algorithms to promote fairness in society. Let us consider the recidivism prediction algorithm for example. What is the purpose of developing and using this algorithm anyway? Is it to put people in jail for more years, or to prevent them from going back to factors that could lead to recidivism (such as poverty, violence, drug and alcohol use) and to help them thrive in society? If the algorithm is reoriented from incarceration to rehabilitation, how would its risk rating change? How can and should the algorithm be redesigned to oppose the mass incarceration of poor people of color and the systemic racial-economic inequality?

Some researchers might argue that these questions are beyond the scope and capacity of AI fairness research. They might say that their goal is to de-bias AI algorithms and create algorithms that are more sensitive to multiple attributes, not to subvert the intersecting system of social injustice in general. I agree that eliminating oppression is not solely the responsibility computer scientists. However, given the wide-reaching effects of biased algorithms that go beyond the academic field of computer science—which span from making individuals serve longer sentences, not get hired, and be denied loans to perpetuating the marginalization of underrepresented groups—algorithms should be developed in a more careful, socially responsible manner. I expect that many AI fairness researchers would actually agree with Noble when she says that the goal of her project is to "eliminate social injustice and change the ways in which people are oppressed with the aid of allegedly neutral technologies," [42, p. 3] because their research, too, has been motivated by the acknowledgment that AI technologies are not neutral but reproduce social injustice.

Moreover, I do not ask AI researchers to bear all burdens. To find the best ways to use AI technologies to resist oppression, AI fairness research should involve *collaborative* projects. Collaboration is to take place not only across disciplines in the academy, but also between communities and researchers. To make recidivism prediction algorithms "fairer" in the strong sense, researchers should have extensive discussions with communities and stakeholders (for example, defendants, prisoners, advocates, law enforcement officers, social workers, judges, and lawyers), rather than making and testing the algorithm only in the lab and then just "throw it in the wild." This discussion-based research is essentially cross-disciplinary: it takes place not only within the field of computer science but across engineering, the humanities, and social sciences. Computer science can benefit from the principles and practices of community-based participatory research (CBPR), philosophical discussions of what fairness is, feminist and critical race studies' emphasis that intersectionality is less about identity but more about power, and in the case at hand, criminology, legal studies, and sociology. Through collaboration across communities and across disciplines, AI fairness research could better find ways to use algorithms to improve fairness and justice in society, as opposed to perpetuating the status quo injustice.

ACKNOWLEDGMENTS

An earlier version of this article was presented at the Feminism, Social Justice, and AI Workshop hosted by *Feminist Philosophy Quarterly* in July 2021. I am grateful to the workshop participants and organizers, whose feedback has been vital to the development of this article. I appreciate very helpful and detailed suggestions from the anonymous reviewers at FAccT. I declare no additional sources of funding, and no financial interests.

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. *ProPublica.* (May 23, 2016). Retrieved May 8, 2022 from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[2] Kwame Anthony Appiah. 2020. The Case for Capitalizing the *B* in Black. *The Atlantic.* (June 18, 2020). Retrieved May 8, 2022 from https://www.theatlantic.com/ideas/archive/2020/06/time-to-capitalize-blackand-white/613159/

[3] David Bauder. 2020. AP says it will capitalize Black but not white. *Associated Press.* (July 20, 2020). Retrieved May 8, 2022 from https://apnews.com/article/entertainment-cultures-race-and-ethnicity-us-news-ap-top-news-7e36c00c5af0436abc09e051261fff1f

[4] Ruha Benjamin. 2019. *Race after technology: abolitionist tools for the New Jim Code.* Polity Press, Cambridge, UK; Medford, MA.

[5] Reuben Binns. 2020. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.* 514-524.

[6] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 2018 Conference on Fairness, Accountability and Transparency. PMLR* 81, 77-91.

[7] Anna Carastathis. 2016. *Intersectionality: origins, contestations, horizons.* University of Nebraska Press, Lincoln, NE.

[8] Sumi Cho, Kimberlé Williams Crenshaw, and Leslie McCall. 2013. Toward a field of intersectionality studies: Theory, applications, and praxis. *Signs* 38, 4, 785-810.

[9] Sarah Ciston. 2019. Imagining Intersectional AI. In *Proceedings of the Conference on Computation, Communication, Aesthetics & X*, 39-48.

[10] Patricia Hill Collins. 2015. Intersectionality's definitional dilemmas. *Annual Review of Sociology* 41, 1-20.

[11] Patricia Hill Collins and Sirma Bilge. 2016. *Intersectionality.* Polity Press, Cambridge, UK; Malden, MA.

[12] Combahee River Collective. 2000. The Combahee River Statement [1977].In *Home Girls: A Black Feminist Anthology*, Barbara Smith (Ed.). Rutgers University Press, New Brunswick, 264-274.

[13] Sasha Costanza-Chock. 2018. Design justice, AI, and escape from the matrix of domination. *Journal of Design and Science.* DOI: https://doi.org/10.21428/96c8d426

[14] Kimberlé Crenshaw. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum*, 139-167.

[15] Kimberlé Crenshaw. 1991. Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color. *Stanford Law Review* 43, 6, 1241-1299.

[16] Catherine D'Ignazio and Lauren F. Klein. 2020. *Data Feminism.* The MIT Press, Cambridge, MA.

[17] Ann Caroline Danielsen and Nicole E Noll. 2020. Communicating about COVID-19 and Sex Disparities: A Guide for Media, Scientists, Public Health Officials, and Educators. *GenderSci Blog.* (June 24, 2020). Retrieved May 8, 2022 from https://www.genderscilab.org/blog/covid-communication

[18] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters.* (October 10, 2018). Retrieved May 8, 2022 from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

[19] Ronald Dworkin. 1981. What is equality? Part 1: Equality of welfare. *Philosophy & Public Affairs*, 185-246.

[20] Nancy Ehrenreich. 2002. Subordination and symbiosis: Mechanisms of mutual support between subordinating systems. *UMKC L. Rev.* 71, 251-324.

[21] Virginia Eubanks. 2018. *Automating inequality: how high-tech tools profile, police, and punish the poor* (1st. ed.). St. Martin's Press, New York, NY.

[22] Eve L. Ewing. 2020. I'm a Black Scholar Who Studies Race. Here's Why I Capitalize 'White'. *ZORA.* (July 1, 2020). Retrieved May 8, 2022 from https://zora.medium.com/im-a-black-scholar-who-studies-race-here-s-why-i-capitalize-white-f94883aa2dd3

[23] 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)*. 1918-1921.

[24] James R. Foulds and Shimei Pan. 2020. Are Parity-Based Notions of AI Fairness Desirable? *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering* 43, 4, 51-73.

[25] Katherine Gasdaglis and Alex Madva. 2020. Intersectionality as a regulative ideal. *Ergo* 6, 44, 1287-1330.

[26] Kathryn T. Gines. 2011. Being a Black Woman Philosopher: Reflections on Founding the Collegium of Black Women Philosophers. *Hypatia* 26, 2, 429-437.

[27] Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. 2018. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning. PMLR* 80, 1939-1948.

[28] Anna Lauren Hoffmann. 2019. Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society* 22, 7, 900-915.

[29] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1-16.

[30] Carolyn Dicey Jennings, Regino Fronda, M.A. Hunter, Z.A. Johnson King, A.C. Spivey, and Sharai Wilson. 2019. *Academic Placement and Data Analysis Report on Diversity and Inclusiveness*. American Philosophical Association.

[31] Zhongjun Jin, Mengjing Xu, Chenkai Sun, Abolfazl Asudeh, and HV Jagadish. 2020. MithraCoverage: A system for investigating population bias for intersectional fairness. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*. 2721-2724.

[32] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning. PMLR* 80, 2564-2572.

[33] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2019. An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*. 100-109.

[34] Michael P. Kim, Amirata Ghorbani, and James Zou. 2019. Multiaccuracy: Blackbox post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 247-254.

[35] Brenda Medina and Thomas Frank. 2019. TSA Agents Say They're Not Discriminating Against Black Women, But Their Body Scanners Might Be. *ProPublica*. (April 17, 2019). Retrieved May 8, 2022 from https://www.propublica.org/article/tsa-not-discriminating-against-black-women-but-their-body-scanners-might-be

[36] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. *ACM Computing Surveys* 54, 6, 1-35.

[37] Milagros Miceli, Julian Posada, and Tianling Yang. 2022. Studying Up Machine Learning Data: Why Talk About Bias When We Mean Power? In *Proceedings of the ACM on Human-Computer Interaction*. Article 34.

[38] Jennifer Miller. 2020. Is an Algorithm Less Racist Than a Loan Officer? *New York Times*. (September 18, 2020). Retrieved May 8, 2022 from https://www.nytimes.com/2020/09/18/business/digital-mortgages.html

[39] Michele Moody-Adams. 2003. Racism. In *Blackwell Companion to Applied Ethics*, R. G. Frey and C. H. Wellman (Eds.). Blackwell, Malden, 89-101.

[40] Giulio Morina, Viktoriia Oliinyk, Julian Waton, Ines Marusic, and Konstantinos Georgatzis. 2019. Auditing and Achieving Intersectional Fairness in Classification Problems. arXiv:1911.01468. Retrieved from https://arxiv.org/abs/1911.01468

[41] Arvind Narayanan. 2018. 21 fairness definitions and their politics. Video. In *2018 Conference on Fairness, Accountability, and Transparency*. (February 23-24, 2018). Retrieved May 8, 2022 from https://youtu.be/jIXIuYdnyyk

[42] Safiya Umoja Noble. 2018. *Algorithms of oppression: how search engines reinforce racism*. New York University Press, New York, NY.

[43] Safiya Umoja Noble and Brendesha M. Tynes (Eds.). 2016. *The intersectional Internet: race, sex, class and culture online*. Peter Lang Publishing, Inc., New York, NY.

[44] Cathy O'Neil. 2016. Weapons of math destruction: how big data increases inequality and threatens democracy. Crown, New York.

[45] Cathy O'Neil. 2017. The era of blind faith in big data must end. Video. *TED Conferences*. (April 2017). Retrieved May 8, 2022 from https://www.ted.com/talks/cathy_o_neil_the_era_of_blind_faith_in_big_data_must_end

[46] Vimal Patel. 2016. Diversifying a Discipline. *The Chronicle of Higher Education*. (March 27, 2016). Retrieved May 8, 2022 from https://www.chronicle.com/article/diversifying-a-discipline/

[47] John Rawls. 1971. *A Theory of Justice* (Original ed.). Belknap Press of Harvard University Press, Cambridge, MA.

[48] Andrew D. Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*. 59-68.

[49] Hua Sun and Lei Gao. 2019. Lending practices to same-sex borrowers. *Proceedings of the National Academy of Sciences* 116, 19, 9293-9302.

[50] Yi Chern Tan and L. Elisa Celis. 2019. Assessing social and intersectional biases in contextualized word representations. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. 13230–13241.

[51] Nenad Tomasev, Kevin R McKee, Jackie Kay, and Shakir Mohamed. 2021. Fairness for unobserved characteristics: Insights from technological impacts on queer communities. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 254-265.

[52] Barbara Tomlinson. 2013. To tell the truth and not get trapped: Desire, distance, and intersectionality at the scene of argument. *Signs* 38, 4, 993-1017.

[53] U.S. Consumer Financial Protection Bureau. 2018. *Data Point: 2017 Mortgage Market Activity and Trends*.

[54] U.S. Equal Employment Opportunity Commission. 1978. *Uniform guidelines on employee selection procedures*.

[55] Sari M. van Anders. 2015. Beyond Sexual Orientation: Integrating Gender/Sex and Diverse Sexualities via Sexual Configurations Theory. *Archives of Sexual Behavior* 44, 5, 1177-1213.

[56] Forest Yang, Moustapha Cisse, and Sanmi Koyejo. 2020. Fairness with Overlapping Groups. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 12 pages.

[57] Ke Yang, Joshua R Loftus, and Julia Stoyanovich. 2021. Causal intersectionality for fair ranking. In *2nd Symposium on Foundations of Responsible Computing. Leibniz International Proceedings in Informatics* 192, Article 7, 20 pages.

[58] Iris Marion Young. 1990. *Justice and the Politics of Difference*. Princeton University Press, Princeton, NJ.

[59] Zhe Zhang and Daniel B. Neill. 2017. Identifying significant predictive bias in classifiers. arXiv:1611.0829. Retrieved from https://arxiv.org/abs/1611.08292