Algorithmic Fairness From the Perspective of Legal Anti-discrimination Principles

Vijay Keswani[1] **and L. Elisa Celis**[2]

[1]Duke University
[2]Yale University
vijay.keswani@duke.edu, elisa.celis@yale.edu

Abstract

Real-world applications of machine learning (ML) predictive algorithms often propagate negative stereotypes and social biases against marginalized groups. In response, the field of fair machine learning has proposed technical solutions for various settings that aim to correct the biases in algorithmic predictions. These solutions remove the dependence of the final prediction on the protected attributes (like gender or race) and/or ensure that prediction performance is similar across demographic groups. Yet, recent studies assessing the impact of these solutions in practice demonstrate their ineffectiveness in tackling real-world inequalities. Given this lack of real-world success, it is essential to take a step back and question the design motivations of algorithmic fairness interventions.

We use popular legal anti-discriminatory principles, specifically anti-classification and anti-subordination principles, to study the motivations of fairness interventions and their applications. The anti-classification principle suggests addressing discrimination by ensuring that decision processes and outcomes are independent of the protected attributes of individuals. The anti-subordination principle, on the other hand, argues that decision-making policies can provide equal protection to all only by actively tackling societal hierarchies that enable structural discrimination, even if that requires using protected attributes to address historical inequalities. Through a survey of the fairness mechanisms and applications, we assess different components of fair ML approaches from the perspective of these principles. We argue that the observed shortcomings of fair ML algorithms are similar to the failures of anti-classification policies and constitute violations of the anti-subordination principle. Correspondingly, we propose guidelines for algorithmic fairness interventions to adhere to the anti-subordination principle. In doing so, we hope to bridge critical concepts between legal frameworks for non-discrimination and fairness in machine learning.

1 Introduction

Anti-subordination and anti-classification are legal principles that form the basis of the current US anti-discrimination laws and most of the academic literature around it (Balkin and Siegel 2003). These principles are fundamental to the development of decision-making policies that counter individual and institutional discrimination. Institutional discrimination, in particular, in areas like employment (Brief et al. 2000; Schilt 2011), education (Barber et al. 2020), healthcare (Chen, Szolovits, and Ghassemi 2019), has led to continued disenfranchisement of marginalized groups and a denial of equal citizenship. Alarmingly, years of such discrimination in various societal institutions, on the grounds of race, gender, sexual orientation, religious identity, or age, have influenced and corrupted public data sources with implicit and explicit biases (Jo and Gebru 2020). Decision-making frameworks that continue to use data sources affected by such systemic biases propagate the disenfranchisement of minority groups. To address these biases, anti-classification argues for the creation of decision-making policies that do not take protected attributes of individuals (like race or gender) into account. When simply ensuring equal treatment is insufficient in addressing the impact of institutionalized discrimination, the anti-subordination principle suggests designing decision-making policies that actively address systemic biases, e.g., through affirmative action policies.

Fair machine learning focuses on similar ideas of designing algorithmic methods to learn automated models that are "unbiased" with respect to the protected attributes of individuals. In the context of data on humans, classification involves implementing an automated policy that can assign individuals to class labels for a specific task using their task-relevant and demographic attributes; e.g., predicting students' exam scores using their past performances or predicting whether a loan applicant would default if given the loan using the applicant's credit history. This policy is technically inferred using a labeled training dataset that contains several pairs of individual attributes and their true class labels. By determining relevant patterns from the training dataset, learning algorithms construct classifiers that imitate the relationship between attributes and class labels in the dataset. If this dataset is sufficiently robust, the constructed classifier can "predict" the class labels for future individuals. However, a large amount of literature has pointed out social biases and negative stereotypes in training datasets (Mehrabi et al. 2021). The classifiers trained using these datasets simulate an inaccurate relationship between individuals' attributes and class labels resulting in reduced performance for the groups that the dataset misrepresents. Even beyond training datasets, model misspecifications negatively affect the performance of classifiers for disadvantaged groups (Mehrabi et al. 2021). Hence, it is often nec-

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

essary to design interventions to ensure that the trained classification models do not propagate social biases. One popular algorithmic way this is currently achieved is by constructing classifiers which have similar performance across all groups and which satisfy certain *statistical group fairness* properties. Popular examples of desired fairness properties include statistical parity (equal selection rate across all relevant groups), equalized odds (all relevant groups should have equal group-specific false positive and true positive rates), and so on (Narayanan 2018; Verma and Rubin 2018). In this regard, one of the goals of the field of fair machine learning is to propose frameworks to construct *fair classifiers* that satisfy (one or more) fairness properties (Barocas, Hardt, and Narayanan 2023).

While fair machine learning algorithms aim to address discrimination in automated decision-making pipelines, recent research has shown that applications of trained fair classifiers may not satisfy the desired goal of ensuring that automated decisions are non-discriminatory. As we discuss in the following sections, either due to an inappropriate choice of fairness measure or an improper selection of parameters in the optimization program, fair classifiers can still have the effect of subordinating marginalized groups. Furthermore, in many high-stakes applications like predictive policing or risk assessment, the use of fairness constraints can have the effect of legitimizing unaddressed systemic issues associated with the use of predictive tools by not actively addressing the structural biases that enable subordination. To understand the reasons for these failures of fair classifiers, we study the design of these systems from the perspective of anti-classification and anti-subordination. We forward the proposition that current popular methods to achieve algorithmic fairness essentially aim to follow the anti-classification principle. Correspondingly, we highlight that the failures of fair classification in practice mirror the failures of anti-classification that have been extensively documented and studied in legal and sociological scholarship. In effect, these failures also constitute a violation of the anti-subordination principle and we emphasize the importance of satisfying this principle to construct effective fair classifiers.

The outline for the rest of the paper is the following. We start with an introduction to anti-subordination and anti-classification principles, discussing their advantages and disadvantages (Section 2). Then, we briefly summarize different modules of fair classifiers and discuss why popular fair classification approaches follow the anti-classification principle (Section 3.1). Next, we categorize various failures of fair classification in practice and relate them to the failures of decision-making policies that follow the anti-classification principle (Section 3.2). The specific failure categories we highlight are (a) focus on *technical or superficial neutrality*, (b) limited focus on the dynamic impact of fairness constraints, and (c) legitimization of problematic applications. Finally, we explore the role of the anti-subordination principle in algorithmic fairness and how the above failures constitute a violation of this principle (Section 3.3). While anti-subordination helps uncover the shortcomings of current fairness interventions, we argue that it can also be used to characterize the design principles for interventions that are relatively more effective in addressing algorithmic discrimination. We correspondingly discuss these advantages of satisfying the anti-subordination principle and address the question of whether this principle can ever be satisfied algorithmically within computational frameworks.

Our primary contribution is the use of the conceptual legal frameworks of anti-subordination and anti-classification to analyze the motivations of current fair classification algorithms. We extend the recent line of work that evaluates the role of technology in reinforcing existing oppressive power dynamics. Eubanks (2018) discusses how social and economic inequalities can be entrenched by biased technologies utilized in government-assisted social services. Mohamed, Png, and Isaac (2020) study the advances and biases in AI from the perspective of the decolonial theory to develop guidelines that can help AI practitioners responsibly deploy their tools. Like our paper, Green (2019) questions the utility of incremental approaches to address automated biases and calls for a rigorous study of the relationship between fairness interventions and their social impact. We review the various modules that compose this relationship from an anti-subordination perspective. Mayson (2018) similarly contends that redressing racial disparity in risk assessment requires foundational changes to the criminal justice system, and cannot be achieved solely through algorithmic changes. Other scholars have also called for a broader analysis of fairness definitions and interventions (Corbett-Davies and Goel 2018; Barocas, Hardt, and Narayanan 2023; Edenberg and Wood 2023). Fazelpour and Lipton (2020) argue for the need to account for the underlying mechanisms that cause discrimination. Alikhademi et al. (2021) and Raghavan et al. (2020) discuss whether fairness interventions can legitimize the problems related to the use of automation in areas like policing and recruitment. Our work explores these issues with applications of fairness interventions using anti-discriminatory principles. It is also important to note that we use legal principles to investigate fair ML designs, and do not confront the question of the legality of fairness interventions, a topic where these principles also feature prominently (Bent 2019; Bornstein 2018).

2 Anti-discriminatory Legal Principles

Legal recourse to address discrimination has been a key part of the civil rights movement in the US. Active legal and political support to address historical prejudices has led to several anti-discriminatory laws and policies. At the same time, legal scholarship on this topic has forwarded certain anti-discriminatory principles to codify the methodology for developing anti-discriminatory policies in different parts of society. The anti-classification and anti-subordination principles are two examples and present popular but distinct paths toward developing anti-discriminatory policies.

2.1 Anti-classification Principle

One of the ways to construct non-discriminatory policies is to ensure that the decision-making policies do not take the protected attributes of the individuals into account when making decisions (Balkin and Siegel 2003). This approach

captures the anti-classification principle; enforcement of this principle prohibits the use of protected attributes in the decision-making process and/or disparity in outcomes across groups when group membership is considered normatively immaterial to the task in question. In other words, the anti-classification principle aims to abolish discrimination individuals face due to illegitimate disparities in treatment or outcome based on their protected attributes.

The (implicit and explicit) use of the anti-classification principle led to the legal elimination of several discriminatory practices. By arguing for equal treatment, anti-classification has been hugely successful in mitigating discrimination in education (Brown v. Board of Ed. 1954), employment (Unlawful employment practices 1964), housing (The Fair Housing Act 1968), and lending applications (ECOA 1974). In settings where minority groups have historically been denied equal access, guaranteeing that decision-makers do not discriminate using protected attributes is a major step toward equitable treatment.

The main advantage of the anti-classification principle is that it is theoretically straightforward to understand and implement.[1] If individuals are unfairly denied access to a particular service due to their race or gender (or their proxies), then an obvious step to address this discrimination is to prohibit the use of group membership when making these decisions. Anti-classification principle has also traditionally been considered more *political palatable*. Policies that claim to not differentiate based on any protected attribute and make decisions regarding individuals based on their "merit" do enjoy wide support in many countries.[2]

Despite the advantages of simplicity and political palatability, interventions based on the anti-classification principle suffer from many critical shortcomings. We discuss these shortcomings below and highlight their impact on addressing discrimination in practice.

(1) Facial neutrality. As mentioned above, policies based on anti-classification are palatable partly due to the simplistic nature of the proposed intervention - "don't see color" after all makes for a catchy slogan against race discrimination. However, these policies risk ignoring the privilege that has been afforded to the advantaged groups across generations. Correspondingly, they often fail to address the structural disparities in socioeconomic attributes across groups that have been created due to discrimination in various decision-making settings across generations.

The push for neutrality of the anti-classification kind has been a major feature of the affirmative action debate in the US.[3] With quotas considered unconstitutional in the US (Knowlton 1978), policymakers have had to use a variety of "indirect" affirmative action methods to achieve the goal of improving minority representation. For example, university administrations in Texas and California (where using race in the selection process is prohibited) base their affirmative action policies on guaranteeing appropriate representations from all different regions of the states (Joshi 2019). Even though these policies aim to improve the racial diversity of incoming university classes, the overall selection process is expected to be racially neutral - exhibiting a strange form of facial neutrality to race while implicitly hoping to address the issue of lack of racial minority representation (Kennedy 2015). Perhaps, at least partially due to this conflict in the internal goal of improving diversity and the explicit focus on neutrality, indirect affirmative policies can be ineffective in practice (Joshi 2019; Bleemer 2023).[4]

Indirect policies follow the "colorblind" definition of anti-classification (Nurse 2014), providing an easily advertisable solution to the vastly complex problem of discrimination, at the cost of limited effectiveness in addressing the underlying structural issues.

(2) Limited focus on dynamic impact of anti-discriminatory policies. In settings where discrimination is isolated to individual-level decisions (i.e., where certain problematic decision-makers discriminate against minority groups), anti-classification can serve as a useful tool to ensure equal outcomes. In reality, cases of discrimination are rarely this simple; they do not happen in isolation (e.g., through laws that implicitly reinforce the subordination of minority groups) and acts of discrimination reinforce disparities across time and context (Gaskin, Headen Jr, and White-Means 2004). Anti-classification, however, has a narrow focus on addressing discrimination within the policies used by individual decision-makers and this narrow focus has been an issue in various applications.

Historically, we see this point feature in the debate around the impact of Brown v. Board of Ed. ruling. The elimination of racial segregation in public schools aimed to to spur systemic changes in other fields (Glickstein 1980). However, the ruling and the following policy changes did not completely address racial disparities in other educational sectors. For instance, desegregation in public schools led to the establishment of private schools in the US South, the student demographics of which were predominantly white (Clotfel-

[1] We say theoretically because, operationally, the implementation of the anti-classification principle can take many forms, some of which are relatively harder to implement than others. For example, simply making race unavailable to the decision-maker can be sufficient to achieve anti-classification in some cases (i.e., "colorblind" policies (Siegel 2010)). However, if other race proxies are available, decisions can still cause disparate outcomes if they "learn" the protected attributes using the proxy variables (Kilbertus et al. 2017). Addressing these proxies requires additional effort.

[2] Often making the problematic assumption that metrics defining "meritocracy" are unaffected by the social dynamics associated with protected attributes (Holmes 2007).

[3] This is especially true with SFFA v. Harvard case, where the plaintiffs claimed that the ruling was "the beginning of the restoration of the colorblind legal covenant that binds together our multiracial, multi-ethnic nation" (Sloan 2023).

[4] Bleemer (2023) studies the impact of California's Prop 209 policy, which ended race-based affirmative action for UC schools and replaced it with region-based affirmative action (i.e., the top four percent of students in each district of California are guaranteed spots in state schools). Their analysis suggests that this move led to a decrease in degree attainment and post-graduation wages of under-represented minority groups. Similarly, simulations of socioeconomic status-based affirmative action demonstrate that they do not improve minority representation as much as direct race-based affirmative action policies (Reardon et al. 2017).

ter 2004). Desegregation in public schools was also rendered less effective due to continued discriminatory housing practices (Henderson and Browne 2004). Another aftermath of the ruling was that several qualified Black teachers and school administrators, some of whom had also by then established schools for Black children, were dismissed once the schools were integrated (Fultz 2004; Will 2019). These examples point to the necessity of analysis of the scope of anti-discriminatory policies. Anti-classification addresses discrimination narrowly by disregarding protected attributes from decision-making; these policies may ignore or exacerbate inequalities that are entrenched in our societal structures due to historical discrimination.

(3) Legitimization of problematic applications. Anti-classification promotes practices that make decision-making frameworks appear neutral with minimal intervention; this superficially-neutral approach, however, can have the harmful impact of legitimizing practices that maintain structural inequalities. Bonilla-Silva (2006) argues how "colorblind" approaches maintain racial inequalities in policing and various other societal domains, while other scholars similarly contend that this ideology is simply a modern form of racism that allows the conservation of racial status quo (Neville et al. 2013). By promoting neutrality as the solution to discrimination, anti-classification can serve as a tool to legitimize systemic inequalities and can create hurdles on the path to broader interventions.

2.2 Anti-subordination Principle

The ideal goal of non-discriminatory policies and legislation is to dismantle the societal hierarchies that enable and propagate social inequalities. To achieve this goal, policies that encourage preferential coverage/selection of individuals from disadvantaged groups have often been proposed to combat social biases that would have otherwise limited opportunities afforded to these groups.

The anti-subordination principle promotes these policies and suggests ways of employing them in a manner that actively tackles systemic inequalities. Anti-subordination scholars, in the context of the civil rights movement in the US, "*contend that guarantees of equal citizenship cannot be realized under conditions of pervasive social stratification and argue that law should reform institutions and practices that enforce the secondary social status of historically oppressed groups*" (Balkin and Siegel 2003). A selection policy satisfies the anti-subordination principle if it does not have the effect of subordinating any group and does not propagate or enable any existing form of subordination that exists in its environment. Anti-subordination can be realized in many ways; it can be achieved using affirmative action or "positive discrimination" policies that provide explicit opportunities and guaranteed representation for historically marginalized groups (Hasnas 2002), or it can be used to argue against the use of tools that can exacerbate discrimination against minority groups (Colker 1987). Importantly, the anti-subordination perspective treats discrimination as a problem that goes beyond the biases of the decision-maker. It examines the broader systemic abuses of power that have led to the subordination of large portions of our population and requires us to consider the viewpoint of individuals from marginalized groups when constructing our non-discriminatory frameworks (Colker 1987)

In the previous section, we discussed certain shortcomings of an anti-classification approach; using the same examples as before, we next analyze them from an anti-subordination lens. Indirect affirmative action policies (like region- or socioeconomic status-based affirmative action) do not satisfy the anti-subordination principle when they don't actively work towards overcoming the structural inequalities faced by minority group individuals (Colker 1986). Additionally, these policies violate anti-subordination by also creating roadblocks for more direct policies to address discrimination (Joshi 2019). An anti-subordination approach would reject such policies in favor of those that encourage the use of race, gender, and other protected attributes to address the lack of representation or disparity in decision outcomes. In the case of Brown v. Board of Education, legal scholars have noted that the ruling had both anti-classification and anti-subordination arguments (Balkin and Siegel 2003). As mentioned earlier, the anti-classification principle is satisfied here since the ruling banned the use of race to determine admittance to public schools. Nevertheless, the basis of the ruling was also to condemn policies that continue subordination along racial lines and, as argued by Balkin and Siegel (2003), these anti-subordination values continue to feature and motivate legal rulings against discriminatory practices. However, an explicit adherence to anti-subordination could have provided a foundation to address superficially race-neutral practices that still existed after the ruling and that continued to create educational disparities along racial lines (see Strauss (1989); Balkin and Siegel (2003) for further discussion on this point).

Central to the debate between the two principles is the question of *what should be the focus and the effect of the anti-discriminatory intervention?* The anti-classification principle argues for a narrow approach, whereby we eliminate protected attributes from having any impact on our decision-making process or outcomes – the focus is on making sure that the decision-maker does not discriminate based on the protected attributes; the effect is to guarantee equality in treatment and/or outcomes of the decisions independent of the group membership. The anti-subordination principle suggests taking a more comprehensive approach by actively using protected attributes to create *reparative mechanisms* through which we can address systemic discrimination – the focus is on assessing the broader impact of the current decisions in reinforcing structural inequalities and, if possible, getting the decision-maker to boost the opportunities available to marginalized group individuals; the effect is to both improve their outcomes and actively counter structural inequalities. It should be clear that the scope of the anti-subordination principle is wider than that of the anti-classification principle, which makes it harder to implement in practice. Considering that we cannot always comprehensively predict the future impact of any policy on the underlying population, it can even be impossible to completely

achieve anti-subordination. Nevertheless, even as an idealistic goal, the pursuit of anti-subordination can help develop policies that are more effective in tackling discrimination compared to anti-classification (Colker 1986).

Our brief summary of these two principles (that have been more extensively discussed in decades of legal scholarship (Fiss 1976; Balkin and Siegel 2003; Nurse 2014; Joshi 2019)) aims to emphasize that the anti-subordination principle can be more effective and extensive in tackling group-based discrimination than the anti-classification principle. We next show that the lessons learned from legal discussions on these principles are relevant to the field of fair machine learning and fair classification as well.

3 Anti-discriminatory Principles in Fair ML

As mentioned earlier, ML algorithms that replace or assist human decision-making in various societal domains often propagate existing social biases in these domains. Algorithmic fairness methods aim to address these biases computationally (Barocas, Hardt, and Narayanan 2023). The notions of *fairness* or *justice* employed by algorithmic fairness interventions are also primarily derived from legal scholarship and legislation on anti-discrimination (Ajunwa et al. 2016; Binns, Adams-Prassl, and Kelly-Lyth 2023). However, the descriptive foundations of the anti-discriminatory legal rulings or legislative policies (i.e., reasons for using them and if they were effective in practice) are relatively under-discussed in the fairness literature. Without this discussion, assessing the potential impact of algorithmic fairness interventions can be difficult. Our work aims to fill this gap by canvassing the principles that underpin various fairness mechanisms.

3.1 Fair ML and the Focus on Anti-classification

For a given classification task, the approach to constructing an automated classifier is to search for a policy, amongst a large set of possible policies (we will call this the set of *feasible classifiers*), that is expected to have the least error in predicting the class labels for current and future samples. To ensure that the classifier's predictive performance is similar across all groups and/or independent of the protected attributes, one can constrain the set of feasible classifiers to only those that satisfy certain statistical group-fairness properties, e.g. statistical parity (Zafar et al. 2017), equality of opportunity (Hardt, Price, and Srebro 2016), or causal independence (Kusner et al. 2017) These properties are defined using "fairness metrics" that quantify the disparity of the classifier performance/treatment across groups.[5] Fair classification algorithms then aim to efficiently find the classifier that minimizes the prediction error while satisfying the fairness constraints (i.e., constraints on the amount of bias captured by fairness metrics) (Kamishima et al. 2012; Dwork et al. 2012; Kusner et al. 2017; Celis et al. 2019; Zhang, Lemoine, and Mitchell 2018; Agarwal et al. 2018). [6]

This removal of classifiers that are deemed "statistically unfair" is an application of the anti-classification principle, since it amounts to removing policies that either use protected attributes for disparate treatment or cause disparate performance across protected attribute values. Note that our characterization of anti-classification in fair ML is slightly different from other papers that make use of this principle. Certain works consider anti-classification to be satisfied if protected attributes (and their proxies) are not used to make decisions (Corbett-Davies and Goel 2018; Davis, Williams, and Yang 2021). Corbett-Davies and Goel (2018) go on to define *performance parity* as a different goal than anti-classification, since ensuring performance parity often requires different treatment for different groups. However, in this paper, we use anti-classification as an umbrella term to refer to all of the above properties; i.e., not explicitly/implicitly using the protected attribute and/or ensuring similar performance for all groups both fall under the anti-classification principle. This is because underlying all these methods is an individualistic idea of "equality" that argues that decision processes/outcomes should be independent of protected attributes and should only be based on an individual's "merits" that are relevant to the given setting.[7][8]

3.2 Shortcomings of Fair Classification

The anti-classification approach to algorithmic fairness is currently the popular approach to address algorithmic biases in real-world settings (Mahoney, Varshney, and Hind 2020). We next highlight various issues associated with this approach and discuss how algorithmic fairness methods suffer from shortcomings that are similar to those of anti-classification presented in Section 2.1.

Technical Neutrality

Choice of fairness metric. Fairness metrics play an important role in constructing fair classifiers - they provide a statis-

[5] See Barocas, Hardt, and Narayanan (2023) for a survey of statistical group fairness properties. We primarily discuss group or sub-group fairness in this paper, since they are grounded in legal ideas of addressing group-specific discrimination. Individual fairness, on the other hand, demands that "similar" individuals be treated similarly (Fleisher 2021). However, individual fairness notions do not always address the group-specific biases discussed in this paper.

[6] This approach summarizes the in-processing framework to construct fair classifiers. In contrast, pre-processing approaches *debias* the training dataset so that the classifier trained using the debiased dataset is non-discriminatory (Kamiran and Calders 2012; Celis, Keswani, and Vishnoi 2020). Post-processing approaches modify a trained classifier to address the biases in the classifier outcomes (Hardt, Price, and Srebro 2016; Pleiss et al. 2017). All three approaches have their own domains of applicability, however, the underlying goal is the same – to address the treatment/performance disparity of classifiers across protected attributes.

[7] This interpretation is consistent with the legal literature on anti-classification as well as recent legal rulings in the US around equal protection (Eidelson 2019).

[8] The indeterminacy of the anti-classification principle has been extensively discussed in legal literature, with the idea that the exact interpretation of this principle depends on the legal context where it's applied (see (Balkin and Siegel 2003; Epstien 2017) for additional discussion). By using it to define all policies that aim to achieve independence from protected attributes, we partly side-step this lack of clarity.

tical mechanism for differentiating between fair and unfair classifiers. Ensuring fairness or being unbiased with respect to these metrics implies that the decision-maker treats all groups similarly – indeed this form of "technical neutrality" is an advertised feature of algorithmic fairness interventions (Weerts et al. 2024). However, like anti-classification, this kind of neutrality may not completely address algorithmic discrimination in practice.

Liu et al. (2018) show that, despite the use of a classification policy that is fair with respect to statistical rate (i.e., the policy is constrained to ensure equal selection rate for all relevant groups), the disparity in the underlying distributions of the majority and minority groups can remain unchanged. They claim that "*conventional wisdom suggests that fairness criteria promote the long-term well-being of those groups they aim to protect. ... [we] demonstrate that even in a one-step feedback model, common fairness criteria in general do not promote improvement over time*" (Liu et al. 2018). To understand their claim, consider the following example: suppose a bank is assessing loan applications based on applicants' credit scores and aims to have an equal selection rate for all racial groups. This equal selection rate policy, however, can end up accepting loan applications from minority group individuals who are more likely to default, leading to a decrease in the credit scores of these individuals and either no change or an overall decrease in the average credit scores of the group. In this case, the use of fair classification would have no positive impact in addressing the credit distributional disparities across racial groups (Munnell et al. 1996). In fact, the results of Liu et al. (2018) imply that, in certain cases, an equal selection rate policy might exacerbate the credit distribution differences across groups by being over-eager in assigning positive outcomes.

Similarly, in healthcare settings, McCradden et al. (2020) and Pfohl, Foryciarz, and Shah (2021) show that outcome-based fairness metrics (e.g., statistical rate or equalized odds) can ignore the underlying population heterogeneity and mask the health inequalities that lead to disparate treatment. McCradden et al. (2020) state "*solutions of algorithmic fairness ... show risks of relying too heavily on the so called veneer of technical neutrality, which could exacerbate harms to vulnerable groups ... algorithmic fairness has not accounted for complex causal relationships between biological, environmental, and social factors that give rise to differences in medical conditions across protected identities*". In these cases, while the anti-classification principle is satisfied by ensuring performance parity across all groups defined by the protected attribute, the core causal structures enabling discrimination remain unaffected. Like anti-classification, parity-based fairness metrics provide the appearance of facial neutrality without significant long-term benefits for marginalized groups.

Similar results demonstrating the importance of the choice of fairness metrics have also been observed in settings where feature distributions change due to individuals' actions. Strategic classification models a number of real-world scenarios where individuals react rationally to institutional classifiers (Hardt et al. 2016; Kleinberg and Raghavan 2020). For instance, loan applicants improve their likelihood of successful loan application by trying to increase their credit score in different ways (e.g., opening new credit lines), and mock test exams (for SAT, GRE, etc.) aim to simulate students' performance on actual exams and explicitly provide pointers for improvement. However, the impact of classification in these settings can be discriminatory; individuals from marginalized groups often have to pay higher costs to positively update their features than other groups (Milli et al. 2019; Hu, Immorlica, and Vaughan 2019). While algorithmic fairness interventions aim to address the above lack of equal opportunity, it does not have that impact in real-world settings. Estornell et al. (2023) observe that fair classifiers statistically become less fair than unconstrained classifiers when individuals can strategically update their features (here fairness is again measured using a standard metric, such as the difference between the selection rate of minority and majority group). Hence, the application of many common algorithmic debiasing techniques does not completely address the negative impact of social biases, even though they do satisfy the anti-classification principle.

Representation parameter selection and balance. The use of fairness metrics in the search for fair classifiers also requires defining the boundary between acceptable and unacceptable amounts of bias, as quantified by the metrics themselves. An example of this boundary is the popular 80% or four-fifths rule in employment settings; for example, in a recruitment setting, if the selection rate of women is less than four-fifths of the selection rate of men, then the hiring company should justify the "business necessity" for this disparity (Biddle 2017; Rutherglen 1987). There has been significant policy and legal debate around the choice of this boundary, with questions about whether the number should reflect population demographics or whether it should address historical entry barriers for minority groups (McKinley 2008; Bobko and Roth 2004). The choice of such parameters in fair classifiers has featured similar discussions.

Continuing with the example of the statistical rate fairness metric, defined as the difference between the selection rate of the minority group vs. the selection rate of the majority group. An equal selection rate for all groups suggests that fairness can be realized by ensuring equal representation of all groups amongst the selected individuals. However, while this is a positive step towards addressing biases, assuming that this policy provides equal opportunities to individuals from all groups (and, hence, is fair) can be premature.

First of all, note that an equal group selection rate does not imply that all groups have an equal voice among the selected individuals; e.g. when there's a large difference between the sizes of different groups in the underlying population. Significant differences in group representation (independent of the eligible population representation) can lead to unstable power dynamics that favor the over-represented groups or propagation of negative stereotypes against the under-represented groups, often necessitating the need for fixed quotas (Ruiz and Rubio-Marín 2008; Kay, Matuszek, and Munson 2015; Noble 2018).

Secondly, when an equal selection rate for all groups does lead to an equal (or almost equal) number of selected indi-

viduals from different groups, it does not necessarily imply that all groups are provided similar opportunities to succeed. This point can again be demonstrated using strategic settings introduced in the previous section. In these settings, individuals can strategically manipulate their features, in response to a classifier's decision, to obtain better outcomes in the future. Strategic manipulation requires paying a certain cost that is reflective of the effort required to make the updates and, in many settings, minority group individuals pay larger manipulation costs than majority group individuals (Milli et al. 2019; Hu, Immorlica, and Vaughan 2019). Keswani and Celis (2023) evaluated fairness-constrained classifiers in strategic settings. They observed that even fair classifiers that use statistical parity constraints induce relatively larger manipulation costs for minority group individuals. In other words, despite having an equal selection rate for all groups, minority group individuals are afforded relatively fewer opportunities (and pay higher costs) to exercise their recourse options against institutional decisions, compared to majority group individuals. Furthermore, by analyzing the relationship between cost functions and statistical parity metrics, they show that using modified fairness constraints – requiring selecting minority group individuals at a higher rate than majority group individuals– can lead to equalization of manipulation costs across groups. That is, unequal selection rates (favoring the minority group) can reduce the disparity in the average recourse costs across groups.

Equal selection rates would ensure a neutral selection process and satisfy the anti-classification principle. Limitations of the training dataset are addressed using this fairness procedure, but only to the extent that all groups are selected at an equal level in the present and future. However, preferential treatment for marginalized groups can sometimes be necessary to provide overall improved opportunities to these groups in domains historically marked by systemic discrimination. Preferential treatment of this form, nevertheless, would violate the anti-classification principle.

Dynamic Impact of Fair Classifiers

Impact of fair classification on the broader environment. A classification algorithm is often a part of a larger decision-making context. In our discussion on Brown v. Board of Ed. ruling, we saw that while desegregated public schools provided equal access to educational resources for students across racial groups, it still led to exacerbated bias against Black teachers, which impacted the quality of education received by Black students.

Fair classification algorithms can have a similar impact in terms of exacerbating bias from different directions. In their analysis of fair classification, Hu and Chen (2020) show that "*at current, making a system 'more fair' as defined by popular metrics can harm the vulnerable social populations that were ostensibly meant to be served by the imposition of such constraints in the first place*". In particular, they observe that classifiers that ensure fairness using standard fairness metrics (like statistical parity or equalized odds) do not necessarily lead to improved social welfare of the protected demographic groups (defined as the average quality of predictions each group receives from the classifier). The result of Keswani and Celis (2023) regarding fairness in strategic settings, discussed in the previous section, is another example of this phenomenon where the use of fair classification does not benefit individuals from marginalized groups. Dai, Fazelpour, and Lipton (2021) analyze the setting where there are multiple decision-makers and only a certain fraction of them include fairness interventions in their decision-making process; they observe that partial compliance may not lead to fair outcomes for marginalized individuals in the affected population. In all of these contexts, fairness interventions do not have an anti-subordinating effect since they either do not reduce welfare disparity when the broader context is taken into account or when interventions are not judiciously implemented and, as argued by Hoffmann (2019), "*efforts to isolate 'bad data,' 'bad algorithms,' or localized biases of designers and engineers are limited in their ability to address broad social and systemic problems*."

Fairness-accuracy tradeoffs. The choice of *fairness tolerance parameter* is also important when selecting the ideal policy to implement. For instance, consider the goal of constructing a classifier that ensures statistical parity (equal selection rates for all relevant groups). In applications, there is often concern that these kinds of constraints can lead to low test/future accuracy (Menon and Williamson 2018), especially when the classifier that achieves the highest accuracy does not satisfy the fairness constraint. Relaxing this constraint then can lead to a larger set of feasible classifiers and, hence, possibly classifiers that achieve acceptable levels of accuracy while satisfying the relaxed fairness constraint (e.g., using the four-fifths rule suggested for recruitment policies (Biddle 2017; Rutherglen 1987)). This concept of fairness-accuracy tradeoff has been explored in many papers that propose algorithms for fair classification.

However, the inherent framing of this tradeoff can be problematic from an anti-discriminatory perspective. One of the primary reasons for unfairness in classification predictions is the presence of a myriad of biases in the training datasets (van Miltenburg 2016; Weissman and Hasnain-Wynia 2011; Chen, Johansson, and Sontag 2018; Obermeyer et al. 2019). These datasets crucially misrepresent the underlying feature distributions of the marginalized groups, as well as, their probability of success for the classification task in question. Correspondingly, a measurement of accuracy using these biased features and their distributions should also be expected to be biased against marginalized groups as well. In the context of the fairness-accuracy tradeoff, if a classifier that satisfies statistical parity has low accuracy over simulated test partitions, that does not necessarily imply that such a classifier has low true accuracy.

Satisfying statistical fairness properties allows algorithm developers to claim that their algorithms are "unbiased" when making their decisions. However, when implemented or evaluated without regard to the broader decision context, it is difficult to assess whether and how these fairness interventions improve the welfare of marginalized groups. Crucially, the goal of achieving parity with respect to statistical fairness metrics can hinder the search for broader reforms to address discrimination, such as either additional changes

to the optimization process or exploration of non-technical avenues to address the biases.

Legitimization of Problematic Applications
The final shortcoming of anti-classification is that it legitimizes decision-making systems that ideally require a systemic overhaul. Similar issues can arise due to the usage of algorithmic fairness as well. Algorithmic fairness interventions address the issues of just one component in a larger discriminatory decision pipeline and, as argued by Kasy and Abebe (2021), popular algorithmic fairness notions " *legitimize inequalities ... rather than questioning the status quo*." Fairness interventions should not be employed to validate an entire decision pipeline. In applications that either require extensive exploration of alternative automated decision-aid tools or discussion on whether automation can be employed equitably at all; superficial algorithmic attempts in these cases will not ensure anti-subordination. We discuss two relevant examples below.

Predictive policing. The last few decades have seen a renewed surge of demands for police accountability (Walker 2006; Archbold 2022). These demands originate from the frightful power imbalance between the law enforcement agencies and the citizens they are meant to protect and often result in blatant abuse of resources and tools by law enforcement officials (Decker et al. 2019; Moore et al. 2018; Chaney and Robertson 2013; Jon Swaine and Lartey 2015). Automated prediction is another tool that can propagate and exacerbate systemic biases against minority groups in law enforcement (Selbst 2017). Predictive policing tools consist of classification models that aim to predict future criminal activity. However, considering that the datasets that are used to train these models are often biased (Brantingham 2017), inaccurate (Akpinar, De-Arteaga, and Chouldechova 2021), or even incomplete (Klingele, Scott, and Dickey 2010; Siegel 2018), it is not surprising that the trained models inherit these flaws. This has led to ongoing criticisms of these tools in practice and even calls for discontinuing their use in many places (Heaven 2020). Additionally, it is unlikely that incremental algorithmic fairness approaches to "debias" predictive policing tools address the deep systemic biases associated with their use. Nevertheless, fairness interventions allow the developers of predictive policing tools to advertise the tool's "unbiased" nature, as in the case of NYPD's supposedly unbiased predictive policing tool, Patternizr (Griffard 2019).

Criminal risk recidivism. Another related example is the use of automated classification in risk prediction. Recidivism prediction instruments predict the likelihood that a criminal defendant reoffends in the future. For example, the COMPAS tool uses information about the criminal history and demographic attributes of defendants to predict whether they will re-offend within two years when released (Dressel and Farid 2018). Beyond the data collection, fairness measurement, and methodological issues with such a classification system discussed in earlier sections, the features and class label can also contain biases that make the use of automated prediction contentious. For instance, Reisig et al. (2007) observe that racially discriminatory social structures amplify the likelihood of recidivism amongst Black defendants. Furthermore, the accuracy of COMPAS tools is similar to the accuracy of the aggregated decisions of individuals with no criminal justice experience, bringing into doubt the robustness of the features used for these tools (Dressel and Farid 2018). Correspondingly, one question that has been raised is whether algorithmically addressing biases is a feasible approach for this application at all, as simply using automation can amplify biases present in the current framework (Alikhademi et al. 2021).

3.3 Why is the Anti-subordination Principle Important to Fair ML?

An anti-subordination approach situates the impact of a decision-making policy in the broader environment within which the policy is implemented. By taking into account the context in which the proposed fair policies are employed, an anti-subordination approach can motivate design choices that ensure that policy outcomes actively tackle historical and systemic biases. Taking an optimization viewpoint, it expands the space of non-discriminatory solutions to be as broad as possible (going beyond just technical solutions), prioritizing the design motivation of a fair decision-making policy to be the improvement of the overall welfare of marginalized groups. From an anti-subordination perspective, fairness interventions can tackle the algorithmic biases in question only by actively confronting the underlying disparities that manifest in the form of biased classification outcomes. In the previous section, we discussed various settings where popular fair classification techniques have no/negative impact – these are indeed examples of violations of the anti-subordination principle. However, many other works in the field of fair ML do attempt to overcome these drawbacks by developing interventions that work from the perspective of the affected groups and take steps towards achieving anti-subordination. We discuss these works below.

Assessing the dynamic impact of fair classifiers. One of the main arguments against anti-classification-based fairness mechanisms was that they fail to account for their broader and future impact. Modeling the future impact of current decisions is a monumental task and recent work has only been able to construct such models in simplified scenarios. Nevertheless, even simple models can provide valuable insight into whether a fair classifier reduces or promotes inequality. For instance, Zhang, Khalili, and Liu (2020) use simple feedback models to simulate individuals' behaviors to classifiers with group-fairness constraints; their results emphasize the need for synergy between statistical fairness metrics and underlying population dynamics. Similarly, performative prediction algorithms aim to construct classifiers that are stable and optimal with respect to the distributions that the classifier predictions can induce (as a result of strategic actions by relevant agents) (Perdomo et al. 2020; Miller, Perdomo, and Zrnic 2021). The ideas explored in these papers can advance anti-subordination principles of evaluating the broader impact of current decisions and can provide a more comprehensive way of choosing fairness metrics.

In terms of additional algorithmic changes to achieve anti-subordination, certain papers have proposed methods to account for the broader discriminatory context. Mullainathan (2018) first model the underlying social process incorporating equity as part of the welfare function. They also show that not using protected attributes can lead to reduced welfare for all groups. Kallus and Zhou (2018) provide fairness interventions that address the lack of robustness in training data due to past biased selection policies. Zhang et al. (2020) also show that fairness constraints can reduce feature disparities between the groups or exacerbate them depending on the context in which the constraints are employed. They further suggest interventions that can ensure that fairness constraints are primarily employed in a positive manner.

Potential for improved model design and parameter selection. Our earlier discussion in Section 2.1 emphasizes the importance of normatively assessing parameter choices to address representational biases. Achieving anti-subordination via appropriate parameter choices does not require significant changes to the optimization program, rather it mainly requires adjustment to fairness tolerance parameters and the evaluation procedure used. Furthermore, like our discussion on affirmative action, reasonable quotas for under-represented groups can also advance anti-subordination. Specifically in the case of media representation (e.g., Google Search results for various occupations (Celis and Keswani 2020)), population demographics should not limit the choice of representation parameters. Rather, the over-representation of marginalized populations in these data sources can help combat historical representational biases that resulted from these media sources. An anti-subordination perspective favors parameter choices that consider the broader context within which classification is employed, although the exact choices may differ for different applications. Importantly, we believe that algorithm designers, with an anti-subordination perspective in mind, play a crucial role in this discussion by assessing the feasibility of different kinds of minority representation.

Similarly, for fairness-accuracy tradeoffs, to truly assess whether a classifier satisfies certain fairness properties for a specified tolerance parameter (and whether it has an anti-subordinating effect or not), it is important to consider robust methods to measure the actual performance of the classifier. In this direction, multiple recent papers have made important strides in pointing out problems with popular quantification for fairness-accuracy tradeoffs and alternate methods to evaluate such tradeoffs. Dutta et al. (2020) show that classifiers can exhibit different fairness-accuracy tradeoffs when evaluated over the "ideal" underlying dataset from the population compared to over a biased dataset. Wick, Panda, and Tristan (2019) similarly observe reduced friction between common definitions of fairness and performance when the impact of dataset biases is accounted for in the evaluation process. Even in real-world implementations of fair machine learning algorithms, reducing outcomes disparities to achieve equity of predictions has resulted in minimal accuracy loss in many applications (Rodolfa, Lamba, and Ghani 2021; Rodolfa et al. 2020).

Assessment of applications of fair classification. As discussed earlier in the context of predictive policing and risk assessment, an anti-classification approach can potentially legitimize applications that suffer from deeper systemic issues. For these applications, an anti-subordination approach to the use of automated prediction will take into consideration the deeper problems associated with these fields and correspondingly question and even discourage the use of automated prediction tools. The entrenched nature of racial inequalities in domains like policing and criminal justice has been emphasized by various civil rights and academic groups, most of whom call for broader systemic changes (Adler, Picard, and Flood 2019). It is unlikely that fairness interventions (that mainly promote neutrality) will have any effect in addressing the sources of these inequalities unless other structural changes are pursued simultaneously.

Feasibility of anti-subordination principle for the development of fair classifiers. In the sections above, we use the distinction between anti-classification and anti-subordination for investigative purposes. We highlight the issues of anti-classification approaches for algorithmic fairness and contend that current fair classification applications suffer from similar issues. We use anti-subordination as an idealized goal to aim for when designing fairness interventions thinking about interventions to address discrimination, but never truly discuss how to comprehensively satisfy this principle. An obvious follow-up question is *how can we design interventions that comprehensively satisfy the anti-subordination principle?*

This question is difficult to answer, primarily because introducing algorithmic interventions to update classification algorithms in isolation may not address all the contextual biases arising in algorithmic applications. For the examples presented in Section 3.2, we noted that the choice of fairness metrics or parameters determines the subordinating effect of fair classifiers. In many of these cases, with appropriate adjustments, one can actively work towards realizing anti-subordination by properly contextualizing the design and impact of the chosen metrics and parameters. However, the discussion in Section 3.2 also highlighted that the impact of these fairness interventions depends on how they are adopted and implemented in practice; fixing biases from implementation issues is often beyond algorithmic changes. Nevertheless, as noted in Section 3.3, the scope of fair classification in achieving anti-subordination is not completely limited. In certain fairness domains, appropriate technical and non-technical adjustments – e.g., extensive audits (to assess the dynamic impact and relevant fairness-accuracy tradeoffs of chosen interventions) and necessary model updates (e.g., boosting minority representation via parameter changes) – can ensure that the algorithm positively counters past discrimination. However, for domains built upon years of systemic discrimination, any fair classifier can serve as a tool to entrench systemic biases, making it unlikely that anti-subordination can be achieved computationally without broader non-technical and structural changes.

Hence, the anti-subordination principle can motivate fairness intervention designs and applications that ensure that

the chosen interventions are effective in addressing individual and institutional discrimination – however, mechanisms to achieve it comprehensively (via a potential combination of technical and non-technical solutions) remains an open problem. All the examples presented in this paper employ automation in critical settings. Considering the possibility of abuse in these settings, if anti-subordination is indeed the goal, then a transparent data collection process is necessary, and classifiers should be constructed only when the impartiality of the application and the robustness of data/model, is guaranteed. As noted by Virginia Eubanks, on biased algorithms for social welfare, "*if there is to be an alternative, we must build it purposefully, brick by brick and byte by byte*" (Eubanks 2018).

4 Conclusion and Future Work

Using the lens of anti-subordination, we study the designs and applications of fair classification algorithms and argue that, despite algorithmic interventions, fair classifiers can still have a subordinating effect. However, by analyzing the gaps in fair classification tools, we do not aim to suggest that the current research on this topic has not been useful. Rather, our goal is to emphasize the presence of the broader discriminatory environments within which these tools are employed. With appropriate interventions, we can take steps toward anti-subordination. Sections 3.3 highlighted research that proposes broader evaluation methods for fairness interventions, improved design choices, and deeper investigation on the role of debiasing technology in propagating discrimination. These papers show that ideas of anti-subordination have been *implicitly* considered by the research community. Our work emphasizes the importance of using anti-subordination as the *explicit goal* of algorithmic fairness.

Beyond technical solutions, *transparency* and *participation* also help counter algorithmic discrimination. Documentation on technical aspects of ML tools can allow for third-party audits and help build trust in these tools when they are used in practice (Griffard 2019). Participation from disadvantaged communities is similarly paramount in addressing automated biases. Discussions on algorithmic fairness should not be limited to the parties with biased algorithmic tools and the parties proposing fairness interventions. As pointed out by Colker (1987) in the case of anti-discriminatory policies, and examined for algorithmic fairness definitions (Saxena et al. 2019; Harrison et al. 2020), public participation (especially from historically-subordinated groups) in framing and assessing fairness interventions can lead to inclusive frameworks to address structural discrimination.

Finally, other legal principles can also be used to motivate fair ML methods. Bornstein (2018) forward the *anti-stereotyping principle*, which requires that individuals not be held to stereotypes associated with their groups. Siegel (2010) aims to reconcile different non-discriminatory legal policies using the *anti-balkanization principle*, which assesses anti-discriminatory policies based on their threat to social cohesion. Many legal scholars have similarly studied anti-discriminatory principles from a non-US perspective (Pager 2007; Weerts et al. 2023). Our use of anti-classification and anti-subordination is one approach to question the current motivations of fair ML; future research should continue this examination along (and beyond) other legal principles of anti-discrimination.

The arguments we make in this paper to point out the shortcomings of anti-classification-based algorithmic fairness interventions bear similarities with many different works on structural inequalities in various societal domains. Seminal works on the structures of *racialized capitalism* (see Alcoff (2021)) and *postcolonialism* (Young 2016) all highlight the role of oppressive power dynamics in continuing to exploit social inequalities for economic gains. The presence of these structures has also been noted in legal domains, including recent work arguing their subordinating impact on racial minorities (Brito et al. 2022). Similar literature on structural inequalities, from the fields of epistemology, philosophy, and political sciences, are extremely important to the pursuit of algorithms that are free from social biases and inequalities, especially considering the fact that there's already decades-long work in these fields demonstrating the disadvantages of superficial approaches to fairness (Carr 1997; Valli 1995). Yet, designs and applications of fair ML often fail to heed the lessons from this wider literature. Using legal principles, our work attempts to connect this literature to the currently popular computational fairness notions and applications, and hopes to encourage the technical fair ML audience to employ the extensive scholarship cultivated in other academic disciplines to assess the impact and effectiveness of proposed algorithmic fairness interventions.

Acknowledgements

This project is supported in part by NSF Award IIS-2045951. We would like to thank Yuvraj Joshi, Aaron Mendon-Plasek, Walter Sinnott-Armstrong, and the community at the Information Society Project at the Yale Law School and the Yale Institute of Social and Policy Studies for helping workshop this topic and for valuable feedback on various drafts of this paper.

References

Adler, J.; Picard, S.; and Flood, C. 2019. Arguing the algorithm: Pretrial risk assessment and the zealous defender. *Cardozo J. Conflict Resol.*, 21: 581.

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

Ajunwa, I.; Friedler, S.; Scheidegger, C. E.; and Venkatasubramanian, S. 2016. Hiring by algorithm: predicting and preventing disparate impact. *Available at SSRN*.

Akpinar, N.-J.; De-Arteaga, M.; and Chouldechova, A. 2021. The effect of differential victim crime reporting on predictive policing systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 838–849.

Alcoff, L. 2021. Critical Philosophy of Race.

Alikhademi, K.; Drobina, E.; Prioleau, D.; Richardson, B.; Purves, D.; and Gilbert, J. E. 2021. A review of predictive policing from the perspective of fairness. *Artificial Intelligence and Law*, 1–17.

Archbold, C. A. 2022. A Look at Police Accountability Through the Lens of the George Floyd Case. In *Rethinking and Reforming American Policing*, 259–288. Springer.

Balkin, J. M.; and Siegel, R. B. 2003. The American civil rights tradition: Anticlassification or antisubordination. *Issues in Legal Scholarship*, 2(1).

Barber, P. H.; Hayes, T. B.; Johnson, T. L.; and Márquez-Magaña, L. 2020. Systemic racism in higher education. *Science*, 369: 1440 – 1441.

Barocas, S.; Hardt, M.; and Narayanan, A. 2023. *Fairness and machine learning: Limitations and opportunities*. MIT Press.

Bent, J. R. 2019. Is algorithmic affirmative action legal. *Geo. LJ*, 108: 803.

Biddle, D. 2017. *Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing*. Routledge.

Binns, R.; Adams-Prassl, J.; and Kelly-Lyth, A. 2023. Legal taxonomies of machine bias: Revisiting direct discrimination. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1850–1858.

Bleemer, Z. 2023. Affirmative Action and its Race-Neutral Alternatives. *Available at SSRN 4319357*.

Bobko, P.; and Roth, P. L. 2004. The four-fifths rule for assessing adverse impact: An arithmetic, intuitive, and logical analysis of the rule and implications for future research and practice. In *Research in personnel and human resources management*. Emerald Group Publishing Limited.

Bonilla-Silva, E. 2006. *Racism without racists: Color-blind racism and the persistence of racial inequality in the United States*. Rowman & Littlefield Publishers.

Bornstein, S. 2018. Antidiscriminatory algorithms. *Ala. L. Rev.*, 70: 519.

Brantingham, P. J. 2017. The logic of data bias and its impact on place-based predictive policing. *Ohio St. J. Crim. L.*, 15: 473.

Brief; Dietz; Cohen; Pugh; and Vaslow. 2000. Just Doing Business: Modern Racism and Obedience to Authority as Explanations for Employment Discrimination. *Organizational behavior and human decision processes*, 81 1.

Brito, T. L.; Sabbeth, K. A.; Steinberg, J. K.; and Sudeall, L. 2022. Racial capitalism in the civil courts. *Colum. L. Rev.*, 122: 1243.

Brown v. Board of Ed., . 1954. Brown v. Board of Education of Topeka.

Carr, L. G. 1997. *" Colorblind" Racism*. Sage.

Celis, L. E.; Huang, L.; Keswani, V.; and Vishnoi, N. K. 2019. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the conference on fairness, accountability, and transparency*, 319–328.

Celis, L. E.; and Keswani, V. 2020. Implicit diversity in image summarization. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2): 1–28.

Celis, L. E.; Keswani, V.; and Vishnoi, N. 2020. Data preprocessing to mitigate bias: A maximum entropy based approach. In *International Conference on Machine Learning*, 1349–1359. PMLR.

Chaney, C.; and Robertson, R. V. 2013. Racism and police brutality in America. *Journal of African American Studies*, 17(4): 480–505.

Chen, I. Y.; Johansson, F. D.; and Sontag, D. 2018. Why is my classifier discriminatory? In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 3543–3554.

Chen, I. Y.; Szolovits, P.; and Ghassemi, M. 2019. Can AI help reduce disparities in general medical and mental health care? *AMA journal of ethics*, 21(2): 167–179.

Clotfelter, C. T. 2004. Private schools, segregation, and the southern states. *Peabody Journal of Education*, 79(2): 74–97.

Colker, R. 1986. Anti-subordination above all: Sex, race, and equal protection. *NyUL REv.*, 61: 1003.

Colker, R. 1987. The Anti-Subordination Principle: Applications. *Wis. Women's LJ*, 3: 59.

Corbett-Davies, S.; and Goel, S. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*.

Dai, J.; Fazelpour, S.; and Lipton, Z. 2021. Fair machine learning under partial compliance. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 55–65.

Davis, J. L.; Williams, A.; and Yang, M. W. 2021. Algorithmic reparation. *Big Data & Society*, 8(2): 20539517211044808.

Decker, M. R.; Holliday, C. N.; Hameeduddin, Z.; Shah, R.; Miller, J.; Dantzler, J.; and Goodmark, L. 2019. "You do not think of me as a human being": Race and gender inequities intersect to discourage police reporting of violence against women. *Journal of urban health*, 96(5): 772–783.

Dressel, J.; and Farid, H. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1): eaao5580.

Dutta, S.; Wei, D.; Yueksel, H.; Chen, P.-Y.; Liu, S.; and Varshney, K. 2020. Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing. In *International Conference on Machine Learning*, 2803–2813. PMLR.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226.

ECOA, . 1974. The Equal Credit Opportunity Act [ECOA].

Edenberg, E.; and Wood, A. 2023. Disambiguating algorithmic bias: from neutrality to justice. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 691–704.

Eidelson, B. 2019. Respect, Individualism, and Colorblindness. *Yale LJ*, 129: 1600.

Epstien, D. C. 2017. Black and White and Gray All Over: How Anticlassification Theory Can Endorse Race-Based Affirmative Action Policies. *U. Pa. J. Const. L.*, 20: 433.

Estornell, A.; Das, S.; Liu, Y.; and Vorobeychik, Y. 2023. Group-fair classification with strategic agents. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 389–399.

Eubanks, V. 2018. The digital poorhouse. *Harper's Magazine*.

Fazelpour, S.; and Lipton, Z. C. 2020. Algorithmic fairness from a non-ideal perspective. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 57–63.

Fiss, O. M. 1976. Groups and the equal protection clause. *Philosophy & Public Affairs*, 107–177.

Fleisher, W. 2021. What's Fair about Individual Fairness? *AIES 2021*.

Fultz, M. 2004. The displacement of Black educators post-Brown: An overview and analysis. *History of Education Quarterly*, 44(1): 11–45.

Gaskin, D. J.; Headen Jr, A. E.; and White-Means, S. I. 2004. Racial disparities in health and wealth: The effects of slavery and past discrimination. *The review of Black political economy*, 32(3-4): 95–110.

Glickstein, H. A. 1980. The Impact of Brown v. Board of Education and Its Progeny. *Howard LJ*, 23: 51.

Green, B. 2019. Good" isn't good enough. In *Proceedings of the AI for Social Good workshop at NeurIPS*.

Griffard, M. 2019. A Bias-Free Predictive Policing Tool: An Evaluation of the NYPD's Patternizr. *Fordham Urb. LJ*, 47: 43.

Hardt, M.; Megiddo, N.; Papadimitriou, C.; and Wootters, M. 2016. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, 111–122.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of Opportunity in Supervised Learning. In *NIPS*.

Harrison, G.; Hanson, J.; Jacinto, C.; Ramirez, J.; and Ur, B. 2020. An empirical study on the perceived fairness of realistic, imperfect machine learning models. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 392–402.

Hasnas, J. 2002. Equal opportunity, affirmative action, and the anti-discrimination principle: The philosophical basis for the legal prohibition of discrimination. *Fordham L. Rev.*, 71: 423.

Heaven, W. D. 2020. Predictive policing algorithms are racist. They need to be dismantled.

Henderson, W.; and Browne, J. A. 2004. Building Housing and Communities Fifty Years After" Brown v. Board of Education". *Journal of Affordable Housing & Community Development Law*, 437–442.

Hoffmann, A. L. 2019. Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society*, 22(7): 900–915.

Holmes, D. G. 2007. Affirmative reaction: Kennedy, Nixon, King, and the evolution of color-blind rhetoric. *Rhetoric review*, 26(1): 25–41.

Hu, L.; and Chen, Y. 2020. Fair classification and social welfare. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 535–545.

Hu, L.; Immorlica, N.; and Vaughan, J. W. 2019. The disparate effects of strategic manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 259–268.

Jo, E. S.; and Gebru, T. 2020. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 306–316.

Jon Swaine, O. L.; and Lartey, J. 2015. Black Americans killed by police twice as likely to be unarmed as white people.

Joshi, Y. 2019. Racial Indirection. *Law & Society: Private Law - Discrimination Law eJournal*.

Kallus, N.; and Zhou, A. 2018. Residual unfairness in fair machine learning from prejudiced data. In *International Conference on Machine Learning*, 2439–2448. PMLR.

Kamiran, F.; and Calders, T. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1): 1–33.

Kamishima, T.; Akaho, S.; Asoh, H.; and Sakuma, J. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 35–50. Springer.

Kasy, M.; and Abebe, R. 2021. Fairness, equality, and power in algorithmic decision-making. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 576–586.

Kay, M.; Matuszek, C.; and Munson, S. A. 2015. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 3819–3828.

Kennedy, R. 2015. *For discrimination: Race, affirmative action, and the law*. Vintage.

Keswani, V.; and Celis, L. E. 2023. Addressing strategic manipulation disparities in fair classification. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–11.

Kilbertus, N.; Rojas Carulla, M.; Parascandolo, G.; Hardt, M.; Janzing, D.; and Schölkopf, B. 2017. Avoiding discrimination through causal reasoning. *Advances in neural information processing systems*, 30.

Kleinberg, J.; and Raghavan, M. 2020. How Do Classifiers Induce Agents to Invest Effort Strategically? *ACM Transactions on Economics and Computation (TEAC)*, 8(4): 1–23.

Klingele, C.; Scott, M. S.; and Dickey, W. J. 2010. Reimagining criminal justice. *Wis. L. Rev.*, 953.

Knowlton, R. E. 1978. Regents of the University of California v. Bakke. *Ark. L. Rev.*, 32: 499.

Kusner, M. J.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual Fairness. *Advances in Neural Information Processing Systems*, 30.

Liu, L. T.; Dean, S.; Rolf, E.; Simchowitz, M.; and Hardt, M. 2018. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, 3150–3158. PMLR.

Mahoney, T.; Varshney, K.; and Hind, M. 2020. *AI fairness*. O'Reilly Media, Incorporated.

Mayson, S. G. 2018. Bias in, bias out. *YAle lJ*, 128: 2218.

McCradden, M. D.; Joshi, S.; Mazwi, M.; and Anderson, J. A. 2020. Ethical limitations of algorithmic fairness solutions in health care machine learning. *The Lancet Digital Health*, 2(5): e221–e223.

McKinley, S. W. 2008. The Need for Legislative or Judicial Clarity on the Four-Fifths Rule and How Employers in the Sixth Circuit Can Survive the Ambiguity. *Cap. UL Rev.*, 37: 171.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*.

Menon, A. K.; and Williamson, R. C. 2018. The cost of fairness in binary classification. In *Conference on Fairness, accountability and transparency*, 107–118. PMLR.

Miller, J.; Perdomo, J. C.; and Zrnic, T. 2021. Outside the Echo Chamber: Optimizing the Performative Risk. *arXiv preprint arXiv:2102.08570*.

Milli, S.; Miller, J.; Dragan, A. D.; and Hardt, M. 2019. The social cost of strategic classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 230–239.

Mohamed, S.; Png, M.-T.; and Isaac, W. 2020. Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology*, 33(4): 659–684.

Moore, S. E.; Robinson, M. A.; Clayton, D. M.; Adedoyin, A. C.; Boamah, D. A.; Kyere, E.; and Harmon, D. K. 2018. A critical race perspective of police shooting of unarmed black males in the United States: Implications for social work. *Urban Social Work*, 2(1): 33–47.

Mullainathan, S. 2018. Algorithmic fairness and the social welfare function. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, 1–1.

Munnell, A. H.; Tootell, G. M.; Browne, L. E.; and McEneaney, J. 1996. Mortgage lending in Boston: Interpreting HMDA data. *The American Economic Review*, 25–53.

Narayanan, A. 2018. Translation tutorial: 21 fairness definitions and their politics. In *Proc. Conf. Fairness Accountability Transp., New York, USA*, volume 1170.

Neville, H. A.; Awad, G. H.; Brooks, J. E.; Flores, M. P.; and Bluemel, J. 2013. Color-blind racial ideology: Theory, training, and measurement implications in psychology. *American Psychologist*, 68(6): 455.

Noble, S. U. 2018. *Algorithms of oppression*. New York University Press.

Nurse, A. 2014. Anti-subordination in the equal protection clause: A case study. *NYUL Rev.*, 89: 293.

Obermeyer, Z.; Powers, B.; Vogeli, C.; and Mullainathan, S. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*.

Pager, S. A. 2007. Antisubordination of Whom-What India's Answer Tells Us about the Meaning of Equality in Affirmative Action. *UC Davis L. Rev.*, 41: 289.

Perdomo, J.; Zrnic, T.; Mendler-Dünner, C.; and Hardt, M. 2020. Performative prediction. In *International Conference on Machine Learning*, 7599–7609. PMLR.

Pfohl, S. R.; Foryciarz, A.; and Shah, N. H. 2021. An empirical characterization of fair machine learning for clinical risk prediction. *Journal of biomedical informatics*, 113: 103621.

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

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

Reardon, S. F.; Baker, R.; Kasman, M.; Klasik, D.; and Townsend, J. B. 2017. Can socioeconomic status substitute for race in affirmative action college admissions policies? Evidence from a simulation model.

Reisig, M. D.; Bales, W. D.; Hay, C.; and Wang, X. 2007. The effect of racial inequality on black male recidivism. *Justice Quarterly*, 24(3): 408–434.

Rodolfa, K. T.; Lamba, H.; and Ghani, R. 2021. Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy. *Nature Machine Intelligence*, 3(10): 896–904.

Rodolfa, K. T.; Salomon, E.; Haynes, L.; Mendieta, I. H.; Larson, J.; and Ghani, R. 2020. Case study: predictive fairness to reduce misdemeanor recidivism through social service interventions. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 142–153.

Ruiz, B. R.; and Rubio-Marín, R. 2008. The gender of representation: On democracy, equality, and parity. *International Journal of Constitutional Law*, 6(2): 287–316.

Rutherglen, G. 1987. Disparate impact under title VII: an objective theory of discrimination. *Virginia Law Review*, 1297–1345.

Saxena, N. A.; Huang, K.; DeFilippis, E.; Radanovic, G.; Parkes, D. C.; and Liu, Y. 2019. How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 99–106.

Schilt, K. 2011. Just One of the Guys?: Transgender Men and the Persistence of Gender Inequality.

Selbst, A. D. 2017. Disparate impact in big data policing. *Ga. L. Rev.*, 52: 109.

Siegel, E. 2018. How to Fight Bias with Predictive Policing.

Siegel, R. B. 2010. From colorblindness to antibalkanization: An emerging ground of decision in race equality cases. *Yale LJ*, 120: 1278.

Sloan, K. 2023. U.S. Supreme Court's affirmative action ruling a 'headwind' for lawyer diversity, experts say.

Strauss, D. A. 1989. Discriminatory intent and the taming of brown. *The University of Chicago Law Review*, 56(3): 935–1015.

The Fair Housing Act, . 1968. The Fair Housing Act.

Unlawful employment practices, . 1964. 42 U.S. Code § 2000e–2 - Unlawful employment practices.

Valli, L. 1995. The dilemma of race: Learning to be color blind and color conscious. *Journal of Teacher Education*, 46(2): 120–129.

van Miltenburg, C. 2016. Stereotyping and bias in the flickr30k dataset. In *11th workshop on multimodal corpora: computer vision and language processing*.

Verma, S.; and Rubin, J. 2018. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*, 1–7. IEEE.

Walker, S. 2006. Police accountability: Current issues and research needs. In *National Institute of Justice (NIJ) policing research workshop: Planning for the future, Washington, DC*.

Weerts, H.; Xenidis, R.; Tarissan, F.; Olsen, H. P.; and Pechenizkiy, M. 2023. Algorithmic unfairness through the lens of EU non-discrimination law: Or why the law is not a decision tree. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 805–816.

Weerts, H.; Xenidis, R.; Tarissan, F.; Olsen, H. P.; and Pechenizkiy, M. 2024. The Neutrality Fallacy: When Algorithmic Fairness Interventions are (Not) Positive Action. In *ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)*, 11.

Weissman, J. S.; and Hasnain-Wynia, R. 2011. Advancing health care equity through improved data collection. *The New England journal of medicine*, 364(24): 2276–2277.

Wick, M.; Panda, S.; and Tristan, J.-B. 2019. Unlocking fairness: a trade-off revisited. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 8783–8792.

Will, M. 2019. 65 Years After 'Brown v. Board,' Where Are All the Black Educators?

Young, R. J. 2016. *Postcolonialism: An historical introduction*. John Wiley & Sons.

Zafar, M. B.; Valera, I.; Rogriguez, M. G.; and Gummadi, K. P. 2017. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, 962–970. PMLR.

Zhang, B. H.; Lemoine, B.; and Mitchell, M. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335–340.

Zhang, X.; Khalili, M. M.; and Liu, M. 2020. Long-term impacts of fair machine learning. *Ergonomics in Design*, 28(3): 7–11.

Zhang, X.; Tu, R.; Liu, Y.; Liu, M.; Kjellstrom, H.; Zhang, K.; and Zhang, C. 2020. How do fair decisions fare in long-term qualification? *Advances in Neural Information Processing Systems*, 33: 18457–18469.