Feminist Philosophy Quarterly

Volume 8 | Issue 3/4 Article 4

2022

Algorithmic Racial Discrimination: A Social Impact Approach

Alysha Kassam
University of California, Irvine
akassam@uci.edu

Patricia Marino
University of Waterloo
pmarino@uwaterloo.ca

Recommended Citation
Kassam, Alysha, and Patricia Marino. 2022. “Algorithmic Racial Discrimination: A Social Impact Approach.” *Feminist Philosophy Quarterly* 8 (3/4). Article 4.

Algorithmic Racial Discrimination: A Social Impact Approach

Alysha Kassam and Patricia Marino

Abstract

This paper contributes to debates over algorithmic discrimination with particular attention to structural theories of racism and the problem of "proxy discrimination"—discriminatory effects that arise even when an algorithm has no information about socially sensitive characteristics such as race. Structural theories emphasize the ways that unequal power structures contribute to the subordination of marginalized groups: these theories thus understand racism in ways that go beyond individual choices and bad intentions. Our question is, how should a structural understanding of racism and oppression inform our understanding of algorithmic discrimination and its associated norms? Some responses to the problem of proxy discrimination focus on fairness as a form of "parity," aiming to equalize metrics between individuals or groups—looking, for example, for equal rates of accurate and inaccurate predictions between one group and another. We argue that from the perspective of structural theories, fairness-as-parity is inapt in the algorithmic context; instead, we should be considering social impact—whether a use of an algorithm perpetuates or mitigates existing social stratification. Our contribution thus offers a new understanding of what algorithmic racial discrimination is.

Keywords: algorithms, algorithmic bias, racial discrimination, structural theories of oppression, proxy discrimination

This paper contributes to debates over algorithmic discrimination with particular attention to structural theories of racism and the problem of "proxy discrimination"—discriminatory effects that arise even when an algorithm has no information about socially sensitive characteristics such as race. Structural theories emphasize the ways that unequal power structures contribute to the subordination of marginalized groups: these theories thus understand racism in ways that go beyond individual choices and bad intentions (see, e.g., Mills 2003; Lebron 2013). Our question is, how should a structural understanding of racism and oppression inform our understanding of algorithmic discrimination and its associated norms? We argue that from a structural point of view, existing frameworks of algorithmic antidiscrimination based on technical understandings of fairness are insufficient;

instead, we must consider the broader social impact of an algorithm's use and whether that use contributes to or ameliorates racial inequity.

Increasingly, algorithms play a central role in a range of public and private practices: predicting recidivism in criminal justice; determining who should be hired, admitted to university, or granted social welfare benefits; evaluating job performance; and suggesting who should get a loan, see an ad, or pay which insurance rate. In these contexts, it has been noted with increasing urgency that the outcomes associated with these processes can be worse for racialized people, for women, and for people in other marginalized communities. As is often noted, one mechanism central to discussions of algorithmic discrimination is the fact that algorithms pick up on social patterns of inequity and past discrimination in the data they are trained on; the algorithms are then are used in ways that perpetuate those patterns (Barocas and Selbst 2016; Bent 2020; Johnson 2021; Prince and Schwarcz 2020). For example, if a hiring algorithm looks at correlations of résumé details with later success to predict future success, then in contexts of existing discrimination and inequity, the result will downgrade résumés of people in marginalized groups. It is well known that these effects arise even if the algorithm has no direct access to information about group membership: instead, the algorithm finds correlations with seemingly innocuous attributes. That discrimination can arise because of these correlations has been called the "proxy problem" and the result "proxy discrimination" (Johnson 2021) or "unintentional algorithmic discrimination" (Bent 2020; Prince and Schwarcz 2020).

The question of how to conceptualize antidiscrimination in the algorithmic context has received a great deal of attention, and there is disagreement over how to understand the concept and associated norms. While some scholars have focused on exploring how existing antidiscrimination norms are relevant to the algorithmic context (e.g., Kleinberg et al. 2018), others argue that because of the possibility of proxy discrimination, norms of the pre-algorithm context, which often focus on largely intention rather than effects, are insufficient (e.g., Ajunwa 2020; Prince and Schwarcz 2020; Bent 2020). In proposing new normative tools, some frame the problem of confronting algorithmic discrimination as one of improving algorithmic fairness. While there is debate over how to conceptualize what algorithmic fairness is, many proposals focus on fairness as a form of "parity," aiming to equalize metrics between individuals or groups—looking, for example, for equal rates of accurate and inaccurate predictions between one group and another (Hellman 2020; see also Bent 2020; Huq 2019; Johnson 2021).

But we argue that, from the structural point of view, framing the problem of algorithmic racial discrimination as a lack of parity is insufficient: algorithms should be evaluated with respect to their broader social impact and whether their use exacerbates or mitigates racial stratification. From a structural perspective, what is racist is understood not with respect to creating unjustified differential outcomes—

as parity-based characterizations might suggest—but rather in terms of whether an act contributes to, or helps to undermine, the existing, asymmetrical social structures of racial inequity and oppression. As we show first, it's worth noting that structural theories of racism add direct and unifying support to claims that we need new norms for antidiscrimination in the algorithmic context—norms focusing on effects and harms over reasons and intentions. But our main point is that starting from the structural understanding of racism, framing algorithmic discrimination in terms of fairness-as-parity is inapt; instead, we should be considering social impact—how a use of an algorithm perpetuates or mitigates existing social stratification. Our contribution thus offers a new understanding of what algorithmic racial discrimination is.[1]

In section 1, we explore algorithmic racial discrimination through examples of proxy discrimination arising in various contexts. In section 2, we discuss the US antidiscrimination normative landscape before the introduction of algorithms, and we explicate recent work proposing new norms of algorithmic antidiscrimination as fairness in a form of parity. In section 3, we draw on structural theories of racism to argue that while proposals to ameliorate discrimination through fairness-as-parity are useful for addressing some forms of discrimination, they do not help address racism as conceptualized structurally. We propose instead a social-impact approach, which centers on considering how an algorithm's use bears on existing racial stratification, and we examine some implications of such a framing. In section 4, we consider broader implications of adopting a social-impact approach, with particular attention to the relationship between "bias" and "discrimination" and to potential intersections with business ethics.

1. Algorithms and the Problem of Proxy Discrimination

In this section, we explicate proxy discrimination through consideration of examples in the contexts of criminal justice, hiring, and setting insurance premiums. Machine learning comes in various forms, often including a "training" phase followed by a "test phase." Many algorithms are deployed in contexts where the goal is to predict an outcome associated with each person in a group, with the results used in

[1] Partly because structural theories remind us that discrimination and inequity takes different forms in different contexts, we focus here just on racial discrimination in the US context, and we leave extensions to other forms of discrimination for other occasions. Because structural theories of racism and the social-impact approach require attention to the particular social context and how racial stratification functions in that context, our proposal can be seen as a contribution toward applying a nonideal methodology in the context of algorithmic fairness, a shift recently urged by Sina Fazelpour and Zachary Lipton (2020).

decision-making. To be represented by data, this outcome must be specified in terms of a specific "objective function" to be maximized: for example, in hiring, the aim might be to maximize number of years at the company or the eventual salary of the employee. These are specific measurable outcomes that stand in for murkier concepts like "successful employee." In these kinds of cases, the training phase involves the use of data in which the various input factors and the objective function are both known; the program then picks up on the various correlations that exist between the input factors and the objective function. In the testing phase, the algorithm is given input factors only, a prediction is made, and the result is checked against the actual value of the objective function. For example, a hiring algorithm might be trained on data that included input factors like years and place of education, years of previous work experience, number of previous positions, and so on, where the objective function is something like "years of employment at the company." In the testing phase, the algorithm would be given the input information only, a prediction would be made for each person for the objective function, and those outputs would be compared with the actual data.

As is often observed, patterns in the training data will emerge in the predictions, and sometimes those patterns are due to existing social injustice and inequality and other social factors we would want to change, not perpetuate. If Black people are less often promoted than white people in a company due to explicit and implicit racism, then algorithms trained on data from that company will predict that Black people will be less successful as future employees. As documented by Safiya Umoja Noble (2018), because people using the internet post images and content in ways that correlate images of Black people with pornography and criminality, algorithms predict that is what people are looking for; as a result, Google answers search queries for "Black girls" with porn, while searching for "white girls" brings up innocuous images. As Gabbrielle Johnson explains, one study found that Google's ad-targeting software resulted in "males [being] shown ads encouraging the seeking of coaching services for high paying jobs more than females" (Johnson 2021, 9944; quoting Datta, Tschantz, and Datta 2015). As Cathy O'Neil (2016) has argued, if an algorithm to predict recidivism takes into account zip codes or "previous family contact with criminal justice system," then the overpolicing of Black neighborhoods and racialized practices like carding or stop-and-frisk mean that even if the algorithm does not include information about race, Black people will be adversely affected.

Consider the much-discussed COMPAS model, which aims to predict recidivism in the criminal justice context. The algorithm is accurate for white and Black people at equal rates but in different ways: it wrongly labels Black people as future criminals at twice the rate of white defendants while labeling white defendants as low risk more often than Black defendants (Huq 2019; Hellman 2020). One mechanism for these outcomes is that algorithms pick up on social patterns of bias and discrimination and then perpetuate them; informally, it is said that the algorithm "bakes in" the existing social injustice. As is often discussed, these effects arise even when the algorithm is not presented with direct information about a person's membership in any particular group; instead, the algorithm finds correlations with seemingly innocuous attributes. The discrimination is thus often unintentional (Prince and Schwarcz 2020; Johnson 2021). For example, "zip code" may seem a like neutral feature, but due partly to the effects of racial stratification, it correlates with race. "Proxy discrimination" refers to the ways that correlations between seemingly neutral attributes and socially sensitive ones resulting from societal inequity and oppression lead to worse results for people in marginalized communities, thus perpetuating inequity.

In our analysis of racism in the algorithmic context, we'll consider a couple of examples beyond the criminal justice context. The first is hiring. In explaining proxy discrimination, Hacker (2018) presents the example of an algorithm that takes into account the distance a potential employee lives from work. It may be that greater distance correlates with worse job performance. But it may also be the case that applicants from farther away are more likely to belong to a particular racial or ethnic group. If both of these are true, then the use of the algorithm will produce proxy discrimination against that group. More generally, if an algorithm uses business success as an objective function, then social discrimination that negatively affects people in marginalized racial groups will produce results that are more likely to predict success for white people and thus to rank them more highly. For instance, the algorithm might pick up on correlations involving features such as Anglo-sounding names and predict future success for people with those features, reflecting existing structures of social inequality and discrimination; this attribute may then correlate with race and ethnicity. As Ifeoma Ajunwa (2020, 1703) explains, algorithms can be used for virtual interviews, tracking facial expression, vocal indications, word choice, and so on, even though facial recognition works less well on darker skin colors. Ajunwa (2020, 1700) also documents the ways that algorithms lead to inequities in how job openings are advertised.

Next, consider the example of setting insurance premiums. In their analysis of proxy discrimination, Anya Prince and Daniel Schwarcz (2020) present a hypothetical insurance algorithm picking up on a correlation between visits to a free website with information on a genetic mutation and the insured person claiming higher future payouts. Plausibly, this could arise because people who have the genetic mutation, and need more expensive health care, visit the website, so there is a correlation between visits and higher claims. If the algorithm assigns to people visiting the website higher premiums on the basis of data showing such correlations, this is proxy discrimination against people with the genetic mutation.

Algorithms are already being widely used in car insurance. As Gert Meyers and Ine Van Hoyweghen (2018) explain, the Dutch insurance company *Fairzekering* (the name is a play on Dutch word for insurance, *verzekering*) has drivers place devices in their cars that track their driving, then uses the data to adjust what a person should pay. The device "will send data every minute such as the location of the vehicle, time, speed, the G-forces the car experiences and what the notifications are from the motor-management" (Meyers and Van Hoyweghen 2018, 428). Based on the data received, an algorithm calculates what premium to charge to reward driving styles thought to correlate with safer driving and thus lower claims. More broadly, in many areas there has been a controversial move toward including information from credit scores in setting insurance premiums.

O'Neil (2016) points out that a driver who must commute to a low-paying job with a chaotic schedule may be regularly driving through what are statistically unsafe neighborhoods; an algorithm tracking that kind of data will charge them higher premiums. Because they are unable to absorb small losses, poor people tend to file more claims; use of the algorithm may thus result in higher claims for poor people. Because Black people in the US tend to be poorer, the burden of such costs would fall disproportionately on them. If there are correlations between proxies for race and higher costs to the insurer, this could potentially be because other drivers are more likely to sue Black drivers.[2] Again, if the algorithm picks up on these correlations and assigns higher premiums to Black drivers, this is proxy discrimination.

2. Equity and Antidiscrimination prior to and in the Algorithmic Context

In this section, we consider the normative antidiscrimination landscape before the introduction of algorithms, briefly survey criticisms of those approaches arising from the possibility of proxy discrimination, then explain the fairness-as-parity conceptualizations crafted specifically to address algorithmic discrimination.

Prior to the use of algorithms, existing norms of antidiscrimination often focused on intentions rather than effects, and it is worth noting here how those norms functioned differently in different contexts. In discussing discrimination, it is useful to distinguish between "disparate treatment" and "disparate impact," a distinction crucial to US law. Legally, disparate treatment involves a practice that is intentionally discriminatory with respect to a protected class such as race, color, religion, sex, or national origin and is illegal. Disparate impact is more subtle and occurs when practices appear to be neutral but result in a disproportionate impact on a protected group. For example, a test for employment might result in people in some class being less likely to be hired. US law here is nuanced: if there is a good reason for the disparate impact, the practice can be legal. For example, if men are more likely to pass

[2] This possibility was suggested to us by Anya Prince and Daniel Schwarcz.

the physical part of the test to become a firefighter because the job requires upper body strength, this can be legal disparate impact. But if there were a physical strength test to become, say, a teacher, and men got hired more than people of other genders, this would be illegal, as physical strength isn't relevant to the job. Furthermore, in addition to ruling out disparate treatment, US law specifies that when disparate impact arises, there must be a "business necessity" for the impact; and that furthermore if the selection rate for one protected group is less than four-fifths of that of the group with the highest selection rate, the employer is at fault (see, e.g., Barocas and Selbst 2016; Raghavan et al. 2020). While the disparate impact clause introduces some nuance, the focus of the law is avoiding disparate treatment and thus intentional discrimination.

In the case of criminal justice and predicting recidivism, matters are obviously more complex, but at a basic level, the relevant US legal norms in play before the introduction of algorithms are found in the Equal Protection Clause of the Fourteenth Amendment. As legal theorist Aziz Huq (2019) explains, this content is focused on two principles: prohibiting the government from classifying people according to categories including race and prohibiting actions that harm individuals due to racial animus or stereotypes. Huq (2019, 1088) calls these "bad classifications" and "bad intent." Again, the focus is largely on intentions and the reasons for which a given decision is made.

In the insurance context, antidiscrimination norms have long been contentious. In the use of "actuarial fairness," widely accepted as an industry standard, premiums should "match as closely as possible" the risk exposure—that is, expected losses—of the insured (Landes 2015). According to this standard, those in riskier positions ought to pay more, and those whose likelihood of loss is less, should pay less, regardless of the cause of the risks (Heath 2007). On the face of it, this entails that some forms of discrimination are actuarially fair: as is often noted, since young men get into more accidents, young men will pay more for insurance even if they are very safe drivers; this is justified under the principle of actuarial fairness. This means if Black people have worse health outcomes than white people because of societal racism and racial inequity, and this leads them to claim more expensive treatments, then actuarial fairness would, in principle, require them to pay more. As these examples suggest, actuarial fairness can also penalize people for things that are beyond their control; for example, a person with a serious health condition, under the concept of actuarial fairness, would pay more for health insurance (see Jha 2012). Relatedly, there is the question of whether there is a causal relationship between the factors and the results: if higher credit scores are correlated with lower payouts, it may seem a higher premium is only justified if there is a causal relationship between credit score and good driving (e.g., an underlying feature of "being careful"); but under the norm of actuarial fairness, showing causation isn't necessary, and correlation is sufficient, if predictive, even if it rests on factors that are not understood (see Gandy 2016). Again, the normative focus in these framings is on avoiding intentional discrimination, not on avoiding disparate impact.

Discrimination by proxy has long existed, but before the use of algorithms, it was often intentional. In the practice of redlining, civic organizations created neighborhood boundaries to influence who should or should not get a mortgage, with racist intentions; this involves using a neutral-seeming proxy—"neighborhood"—to discriminate on the basis of something else—race. The fact that discrimination by proxy used to be typically intentional is reflected in the fact that, as we've seen in the examples, existing legal and ethical analyses focus largely on ruling out disparate treatment and intentional racialized differential treatment.

With respect to the question of whether these norms are sufficient for the algorithmic context, scholars from various disciplines have argued that they are not, and that the unintentional proxy problem is central among reasons that discrimination needs to be reconceptualized (see, e.g., Barocas and Selbst 2016; Chander 2017; Bent 2020; Huq 2019; Prince and Schwarcz 2020). With the use of algorithms, discriminatory and harmful effects can often be unintentional, as we may not know the range of correlations in our data or how the patterns have been influenced by discrimination or bad social attitudes and practices. When it is unintentional, proxy discrimination counts as a form of disparate impact, not disparate treatment.

With respect to harms, notice that because algorithms scale, they can produce disparate impact effects that are more widespread and more pronounced than practices before the introduction of algorithms. Prince and Schwarcz (2020) identify several impacts of proxy discrimination that conflict with the aims of antidiscrimination law.[3] Most importantly for the present context, not only can proxy discrimination "thwart" antisubordination goals, but it can "affirmatively promote the opposite result"—reinforcing legacies of historical discrimination (Prince and Schwarcz 2020, 1296). That is, proxy discrimination can perpetuate and legitimize past inequity: if people with a certain socially sensitive characteristic have been less likely to succeed at work because of discrimination, and an algorithm predicts that people with that feature are less likely to succeed and so determines that they don't get hired, then the use of the algorithm is perpetuating the discrimination. Another impact is that proxy discrimination can undermine efforts at social solidarity.

[3] In their characterization, Prince and Schwarcz (2020) say that proxy discrimination includes only cases where the discrimination benefits the organization deploying the algorithm. This is narrower than our definition, but the difference is not relevant for our argument. In our view, the harms listed apply whether or not the discrimination benefits the discriminator.

Antidiscrimination laws are often aimed at communal sharing of risks and costs that should not be borne by individuals: a person with a genetic anomaly should not be forced into astronomical insurance payments because of something beyond their control, and one aim of health insurance is the social sharing of that burden. But proxy discrimination allows the cost to shift back to the individual. A third impact runs counter to aims preventing stereotyping: when an algorithm discriminates by proxy for people of a certain race or ethnicity, it effectively results in a stereotype; rather than being treated as an individual, a person is regarded as a representative of a group, whether they share that group's statistical profile or not. And because proxy discrimination works in ways that are difficult to predict, it is likely to have chilling effects on the expressive or associational activities of people in protected groups: if a person would be penalized for being treated as a member of a group, they may be less likely to engage in activities connecting them with others in that group.

A further harm is that, as O'Neil (2016) also points out, proxy discrimination can create pernicious feedback loops: a defendant wrongly judged to be a high recidivism risk will be sentenced to a harsher sentence, but the experience of being in prison can lead to more engagement with the criminal justice system, thus creating the impression that the risk factors identified by the algorithm are genuine when in fact the result was created by the effects of the algorithm itself (O'Neil 2016). More generally, Prince and Schwarcz (2020, 1296–97) raise the possibility that proxy discrimination in hiring can lead to lack of steady employment, leading to difficulties getting insurance and health care—which in turn make higher education even less accessible. This type of feedback loop "makes proxy discrimination by AIs particularly pernicious," because "it is the inequitable outcome from one silo that makes the use of that outcome as a proxy rational in the next silo" (Prince and Schwarcz 2020, 1297).

Critics of existing normative frameworks often focus on specific contexts. In the hiring context, Ajunwa (2020) points out that proving illegal discrimination, always difficult, is exceedingly so when algorithms are used: the employee would have to identify a policy or practice that caused the adverse employment outcome, compile relevant statistics to show that the policy has a disparate impact, and rebut the employer's defense that the policy is justified by a business necessity. Given the ways that algorithms work, it is particularly difficult for an applicant or employee to show that a given algorithm caused the discriminatory situation. Jason Bent (2020, 807) points out that in the context of the rule that allows for disparate impacts but only in cases of business necessity and only in restricted ways, the structure of the algorithm itself may provide a business necessity defense to disparate impact liability—as long as the algorithm is predictive, it is predictive of something, and the question then is just whether that something can be shown to be relevant in the right way.[4]

With respect to criminal justice, Huq (2019) argues existing norms such as "bad classifications" and "bad intent" fail to capture the full spectrum of racial issues that can arise in the use of algorithmic tools in criminal justice. Because of the way algorithms rely on historical data, the norms against racial classification may not apply; as the proxy problem shows, eliminating intentional discrimination on the basis of race may leave outcomes unchanged.

With respect to insurance and actuarial fairness, we've discussed that the use of credit scores means that poorer and Black people will pay more for insurance whether or not they are good drivers; O'Neil (2016) says that this approach puts people into "buckets" of people with similar risk profiles and fails to treat them as individuals. We would add that while risk assessment generally uses approaches that consider features, and thus generalize, because of access to detailed data, and because algorithms scale, these aspects and their effects will be more pronounced in the context of algorithms. Prince and Schwarcz (2020) argue that norms in the insurance context should change partly to take into account whether there is a plausible causal story linking an attribute with costlier claims.

In response to the inadequacy of framings based on intentions and disparate impact, some proposals for avoiding algorithmic discrimination focus on improving algorithmic fairness, understood as a form of parity. Computer scientists, legal theorists, philosophers, and others have proposed a wide range of characterizations of what it means to be algorithmically "fair"—by one count, twenty-one different definitions (Huq 2019, 1115). For some of these, there are impossibility results: an algorithm's fairness in one sense is typically incompatible with its fairness in others (see Johnson 2021). Notably, there is agreement that partly because of proxy discrimination, fairness and the amelioration of discrimination cannot be achieved in the algorithmic context through the removal of information about sensitive characteristics. In a discussion of what he calls "algorithmic affirmative action," Bent describes "widespread consensus" among machine learning scholars that algorithmic fairness cannot be accomplished by "hiding protected characteristics" from the algorithm (Bent 2020, 807). In fact, the inclusion of information about socially sensitive characteristics like race is often presented as necessary to improving algorithms, as with this information, we can achieve improved accuracy for people in those groups.

[4] Raghavan et al. (2020, 478) also point out that "machine learning may discover relationships that we do not understand," and they conclude that "a statistically valid assessment may inadvertently leverage ethically problematic correlations."

In an overview of a range of algorithmic fairness proposals, Bent (2020, 817) characterizes "group fairness" as attempting to measure fairness "by comparing the target variable outcomes of a machine-learning process between two groups sorted along the sensitive variable." Group-fairness approaches often appeal to technical definitions that themselves focus on symmetry or parity in the relevant comparison between target variable outcomes (Huq 2019; Bent 2020; Hellman 2020). The simplest form of group fairness is demographic parity, which aims to match proportions in outcomes to proportions in populations. For example, if the applicant pool for a job is 20 percent Black, the outcome should be that 20 percent of people hired are Black. This form of group fairness is sometimes set aside as untenable, on grounds that the result may seem unfair to those predicted to be successful who are not then recommended in the decision process (Bent 2020; see also Pessach and Shmueli 2020)

More subtle parity definitions focus on other measures, such as an equal rate of accuracy among groups, an equal rate of false positives, or a balance of other quantities. The metric used by ProPublica in criticizing the COMPAS algorithm looked at "how frequently false positives are conditional on being in fact a low-risk person" for people in that group (Huq 2019, 1054–55). The metric of "equal predictive value" measures whether scores are "equally predictive of the target trait for members of one group as for members of the other" (Hellman 2020, 827); error-rate balance obtains when people of each group who have or lack the target variable are equally likely to be accurately scored by the test (Hellman 2020, 828).

For example, as Huq (2019, 1115–16) explains, applying fairness metrics in the use of the COMPAS model, we might consider whether equal proportions of different racial groups are deemed likely to reoffend; we might consider whether there is a single numerical cutoff used to demarcate the "likely" from the "unlikely" to reoffend; we might consider whether defendants in different racial groups assigned the same risk score are equally likely to actually recidivate; or we might compare each racial group and "ask how frequently false positives are conditional on being in fact a nonrisky person" for people in that group. The details of these characterizations can be technical and complex. But a vivid example arises in the COMPAS case. Here the creators of the algorithm defended its fairness by pointing out that it was equally likely to be accurate for Black and white defendants. But ProPublica and others argued that its unfairness arises because it is inaccurate for Black and white defendants in different ways: ProPublica charged that the algorithm "was particularly likely to falsely flag black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants" (quoted in Huq 2019, 1048). That is, Black people were more often wrongly classified as likely to recidivate.

Reflecting partly on the COMPAS case, Deborah Hellman (2020, 845) argues that among all the possible metrics for parity, we should focus our attention first on "error ratio parity"—"the ratio between false positive and false negative rates"—for various groups. Here, she says that an algorithm ought to "set the balance between false positives and false negatives in the same way for each group" (845). While balanced ERP is not the only item for attention in her view, it is "suggestive of unfairness" and points toward the need for further investigation and caution.

3. Structural Racism and Algorithmic Discrimination: The Need for a Social-Impact Approach

We've seen in section 2 that characterizations of racial discrimination before the introduction of algorithms often focused on intentions and the reasons for a given decision, and that new proposed frameworks often focus on understanding discrimination in terms of unfairness, to be ameliorated through the pursuit of parity. But in this section, we show that from the point of view of structural theories of racism, these new frameworks are not fully adequate because they fail to consider the ways that even seemingly fair algorithms can perpetuate wider social patterns of stratification. While parity views like Hellman's aptly focus on effects rather than intentions, and will flag some cases of unfairness correctly, parity doesn't account for the structural nature of discrimination. We argue first that adopting a structural theory of racism provides a new, unifying argument to the range of reasons given in different contexts for why focusing on reasons and intentions is inapt in the algorithmic context. But our main point is that structural theories show the need for a normative framework that goes beyond fairness-as-parity to consider "social impact"—how an algorithm's use bears on existing social stratification.

Scholars have long argued that inequity and oppression are better understood as systemic and institutional problems rather than individual ones, and philosophers have drawn on this insight to offer theories of racism that reflect it (see, e.g., Mills 2003; Lebron 2013). Instead of analyzing racism at the individual level, we should see racism in terms of unequal power structures that subordinate those in minority groups. According to Charles Mills (2003), analyzing racism at the individual level makes it seem symmetrical or, in other words, makes it seem that Black people are as culpable as white people when having prejudices. This obfuscates the asymmetrical power dynamics of whites versus minorities (Mills 2003). Instead, racism and other forms of oppression should be viewed as systemic, where the word "systemic" refers to the institutions, policies, or social structures that create disparate impacts for historically marginalized communities. This removes intentionality from the picture, demonstrating how there can be racism even in the absence of racist or otherwise bad intentions, beliefs, or attitudes. Furthermore, such a framing sees racism as applying to specific groups in specific ways: the problem is not the unwarranted drawing of distinctions but rather sustaining an existing racial hierarchy. Structural approaches reject the idea that what makes an action racist is best understood in terms of unjustified differential outcomes. Instead, they emphasize that what makes an action racist or sexist or discriminatory is that it contributes to the hierarchical system; what makes an action antiracist is that it helps undermine or dismantle that system.[5] This development in our understanding of racism helps explain why, despite the decline of overt racism, we nevertheless see marginalized minorities continue to have less wealth and fewer opportunities compared to white counterparts (Mills 2003).

In section 2, we presented some important critiques of norms based on intentions and disparate treatment. But none of these critiques explicitly invokes a structural theory of racism. We claim that a structural point of view provides a new and powerfully unifying argument for moving away from focusing on reasons and intentions and toward considering harms and social impacts. Of course, the general point that structural theories of inequity show the insufficiency of existing frameworks of antidiscrimination centered on intentions and individual reasons is well known. But in the algorithmic context, this point takes on new significance. Proxy discrimination is often unintentional, and effects on marginalized communities transcend disparate treatment and take a variety of forms. Because algorithms scale, the relevant effects come with a greater impact than similar decisions before the use of algorithms. Shifting our focus to a structural one gives a new, direct, and general argument that the effects of using algorithms can be discriminatory or oppressive when their outcomes harm marginalized communities, regardless of intentions. The harms of proxy discrimination are well documented; taking up structural theories of racism helps us understand when and why those harms and social impacts are racially discriminatory.

While this point about unification is significant, our main argument is that structural theories show the need for a normative framework that goes beyond parity definitions of fairness and considers "social impact"—how the use of an algorithm bears on existing racial stratification. Structural understandings of racism and oppression highlight that stratification and power dynamics are asymmetrical across groups, that some groups are dominant while others are subordinated, and that evaluation requires considering the ways that impacts can bear differently on those in dominant versus marginalized communities. It follows that generally what makes a given use of an algorithm discriminatory is whether it contributes to existing

[5] In an analogous analysis in the context of sexism, Marylin Frye (1983, 38) argues that sexism is not about irrelevant markings of the sex distinction but rather about whether and how our actions "create and enforce the elaborate and rigid patterns of sex-marking and sex-announcing which divide the species, along lines of sex, into dominators and subordinates." See also Iris Marion Young's (2011) analysis of oppression.

stratification. Because different groups have different background conditions, starting points, and experiences, effects that are similar from the point of view of metrics fairness will have different impacts on those groups.[6] In a "social impact" characterization, an algorithm is thus racially discriminatory if its use harms marginalized people in ways that add to racial stratification.

While he does not discuss structural theories of inequity, Huq's analysis in the criminal justice context provides an example of a practical approach that would be justified by the normative social-impact framing. In this context, Huq (2019, 1104–5) argues that parity characterizations of algorithmic fairness are inapt because they ignore the ways that the operation of the criminal justice system in the contemporary US generates special harms to those in Black communities, harms that systematically exacerbate racial stratification. Overpolicing, harsher treatment of Black defendants, and the white cultural use of inherent Black criminality to justify social racial discrimination are among an array of mechanisms creating such harms, which have greater negative spillover effects for Black communities than for white ones. False positives—wrongful imprisonment or harsher sentences for those who do not deserve them—perpetuate an existing injustice for Black people in ways that do not apply to members of other racial groups. Error ratio parity, a norm in which we aim for the same proportion of false negatives and false positives for Black and white defendants, can thus lead to outcomes that bear more harshly on those already marginalized. According to Huq (2019, 1113), a proper instantiation of racial equity in the criminal justice context would account for the fact that negative spillover effects are substantially greater for racialized minorities than for the racial majority. Thus evaluation of algorithms should center directly on whether and how their use contributes to racial stratification.

Crucially, since there will be a class of crimes for which a greater benefit will be required to achieve net positive effects when Black suspects are being evaluated by the algorithm, it follows from Huq's proposal that the risk threshold for Black suspects would be set at a higher level than the threshold for white suspects (Huq 2019, 1114). Instead of error ratio parity, we should consider different error ratios depending on the group in question. In essence, the proposal involves considering the impact of the use of the algorithm on people in marginalized groups holistically rather than looking at technical definitions of fairness in the working of the algorithm. Huq's

[6] It may seem that a challenge in considering social impact is that the analysis would require taking into account some information about membership in marginalized groups. As discussed above, however, this challenge is not particular to the social-impact framework. Parity definitions also require that we use information about socially sensitive features, and it's generally accepted that avoiding the proxy problem will require giving algorithms this kind of information (Bent 2020, 816).

approach thus considers the ways algorithms can disproportionately perpetuate harms for marginalized communities, and his framework determines an algorithm's appropriate use based on whether it intensifies or mitigates these harms. This is a basis for decision-making that would ameliorate algorithmic racial discrimination as we are understanding it here; thus, one implication of adopting a social impact approach would be that adopting different risk thresholds for different social groups would be normatively justified.

Next let's explore analogous implications for the two other contexts we've considered—hiring and insurance. In the case of hiring, notice that error ratio parity would invite us to consider whether an algorithm is correct in its predictions in similar ways for people in dominant groups and people in marginalized groups. But if society is such that people in marginalized groups are less likely to be promoted, have higher sales, or achieve the milestones that constitute the objective function, then an algorithm may well be equally predictive for people in various groups and still end in outcomes favoring those in the dominant group. From the structural point of view, if algorithms bear differently on people in different groups, we have some reason to consider using different processes for those groups. A social-impact framework will call our attention to disparate impact in which people in marginalized racial groups are less likely to be hired, and this would entail reason to believe that such a use of an algorithm would be racially discriminatory and therefore wrong.

With respect to further potential implications of this idea, Ajunwa (2020) argues that the difficulty for hires in establishing the relevant facts of wrongful discrimination (e.g., that there has been disparate impact, that it is not a business necessity) should cause us to shift our thinking to what she calls "discrimination per se"—if the use of proxy variables has the potential to result in adverse impact, the burden of proof shifts to the employer to show that its practices are nondiscriminatory. In a hypothetical example that Bent (2020) offers, a city notices that the parts of an exam used in making promotion decisions for firefighters have differential effects in which the written part is more highly predictive of success for white people and the field test is more highly predictive of success for nonwhite people. The city, Bent says, may well have good reasons for weighting the two tests differently for the two groups; among others, these could include reasons to do with avoiding disparate impact from weighting the tests equally.

As we've explained, some characterizations of fairness as parity focus on inaccuracies, comparing false positives, negatives, or ratios. But in our framing, it is striking that even an algorithm with perfect predictions can be used in a way that perpetuates discrimination and inequity. The hiring context showcases this possibility: if an algorithm is trained on data reflecting inequity, and this form of inequity persists in our society, then even an algorithm that is perfect in its predictions can be perpetuating proxy discrimination. Parity definitions of fairness that rest on comparing accuracy rates would fail to capture the inequity of these results. In his discussion of the hypothetical example in section 2, in which distance from work correlates negatively with work performance, Hacker (2018) distinguishes between discrimination resulting from proxy discrimination and that resulting from biased data; he argues that while "getting rid of biased training data generally *increases* predictive accuracy, the elimination of proxy discrimination *reduces* it" (italics in original). But if our society has the same patterns of inequity that produced the data, then plausibly getting rid of biased training data would also reduce predictive accuracy: an algorithm trained on data in which Black employees get promoted less often than white employees will correctly predict that Black employees will get promoted less often than white employees.

With respect to the insurance context, we've seen how the use of algorithms can result in higher premiums for racialized people and how, from the point of view of actuarial fairness, this is not a problem: correlation with predicted payouts is the only normative consideration. As mentioned above, the normative argument against actuarial fairness is partly that it penalizes individuals for their membership in certain groups. Central to this debate have been questions about causation and responsibility: if being a member of the group is causally connected to making claims, because of actions the person is responsible for—say, because the group is "convicted of drunk driving"—then it is said that this may justify a higher premium; if not—say, because the group is "born with a certain gene"—then a higher premium is thought to punish people for what is beyond their control.

Again, from the point of view of a structural approach to algorithmic racism, the relevant considerations concern not causation and responsibility but rather how the impacts relate to existing stratification. On the face of it, the use of information in credit reports would be wrong, on grounds that, as we've seen, poorer people tend to have worse credit scores, racialized people tend to be poorer, and higher premiums make people poorer. In this way of looking at things, actuarial fairness is wrong not (or not only) because of considerations due to causation and responsibility but rather because its pursuit in the context of algorithms will tend to perpetuate existing social inequality. In the insurance context, premiums that create higher costs for Black drivers will perpetuate racial inequity, even if the reasons for those higher premiums are in accordance with error ratio parity; for the reasons we've considered, even fully accurate predictions will result in such higher premiums.

In the social-impact characterization of algorithmic racial discrimination, the focus is directly on effects rather than intentions, and the effects are understood through the lens of how those effects feed, or counter, the social patterns that constitute systemic racism, inequity, and oppression. Therefore, an algorithm that is fair in any sense of parity could still be discriminatory.

4. Further Implications of a Social-Impact Approach

We've seen in section 3 a few implications of adopting a social-impact approach: it would provide normative justification for using different thresholds in criminal justice (as Huq recommends), for weighing criteria differently for different groups in hiring (as in Bent's hypothetical example), and for setting insurance premiums in ways that that conflict with the norm of actuarial fairness. In this section, we consider a few broader implications.

First, adopting a social-impact approach has implications for how we understand the relationships among discrimination, bias, and fairness. It is common in the algorithmic context to talk of "bias": often, the problem is understood as "algorithmic bias" and the fairness-as-parity proposals are developed with the aim of ameliorating that bias. As Birhane and colleagues argue, "fairness" and "bias" in the algorithmic context are often understood abstractly, as linked to "neutrality" or "objectivity" (Birhane et al. 2022, 951). For example, "bias" may be understood to indicate the drawing of distinctions when they are unwarranted or irrelevant in context; bias would thus arise whenever differential outcomes arise for no good reason and could be against any group. Gabbrielle Johnson (2021, 9941) characterizes algorithmic bias generally in terms of "inherit[ed] social patterns reflected in an algorithm's training data." As she says, her broad characterization of bias in general means that virtually all inductive reasoning includes some form of bias, and it includes biases that are "epistemically reliable and morally unproblematic" (Johnson 2021, 9951). Relatedly, O'Neil (2016, ch. 8) points out that generalizing about people from behavioral data—putting them into metaphorical "buckets"—has the problem that it means we are asking "How have people like you behaved in the past" rather than the more apt question "How have you behaved in the past?"

Our analysis results in a characterization of algorithmic discrimination that differs from these understandings of "bias" because it emphasizes the asymmetrical and hierarchical stratification of racism. In the social-impact framing, algorithmic racial discrimination applies to racial groupings asymmetrically: unlike these conceptualizations of bias, which could be against any group, algorithmic racial discrimination is understood as a harm specifically against people marginalized by racism. The resulting concept is thus narrower and more targeted than "bias" in the sense of any unwarranted distinction. In Johnson's framing, our conceptualization of algorithmic discrimination could be a candidate for understanding the particular "problematic" biases she considers a subspecies of bias generally. With respect to the "bucket" metaphor, structural theories of racism emphasize that generalizing about people based on data takes on a particularly pernicious aspect when the groups in question are racially marginalized groups.

This way of understanding algorithmic discrimination has implications for debates over the role of intuitive fairness in characterizing discrimination. In a response to Huq's (2019) proposal to consider different risk thresholds for people in different racial communities, Deborah Hellman (2020) points out that it requires attention to how the use of an algorithm affects a group overall—including those in the group not directly scored by the algorithm. She argues that this scope is inappropriate, on grounds that it leads to an inapt conceptualization of algorithmic fairness: we cannot generally make up for unfairness to some members of a group with a benefit to other members of that group. For example, she says, if the concern is that the use of COMPAS means Black people are treated unfairly compared to white people, we cannot ameliorate this unfairness by a method that provides a benefit to Black people overall and thus possibly to Black people not scored by the algorithm (Hellman 2020, 845).

From our perspective, a response to Hellman would be that any contrast between decision processes fitting our fairness intuitions and those aimed toward reducing discrimination shows not that the latter are misguided but rather that fairness may not be the central value in antiracism. The example she appeals to is framed in terms of how an algorithm affects individuals compared to other individuals and imagines fairness along the lines of applying a relevantly similar criterion in all cases. But as we've seen, structural theories of racism frame antidiscrimination in systemic, asymmetrical terms. The relationship between fairness and antiracism is obviously complex; our point here is just that taking up a structural perspective shows how addressing racism in the algorithmic context may require going beyond fairness.

There are also implications related to business ethics. It might be objected that our framing of discrimination is inapt for contexts of private enterprise where corporations, rather than governments, are making decisions and bearing relevant costs and benefits. Huq, for example, argues that in the criminal justice context, which is public, a bifurcated decision procedure with one threshold for white people and another for Black people has special justification that would not apply to more traditional forms of affirmative action.[7] One reason he gives is that setting different risk thresholds in the criminal justice context could be ideal not only from an equity point of view but also from a social efficiency one: if we are aiming to minimize negative effects of policing overall while maximizing potential benefits, we might most effectively do that by treating different groups differently. And in the context of business and private enterprise, this reasoning would not apply, as decisions are usually based on what's best for the company, not what's best for society.

[7] A further reason he gives for this conclusion is not applicable to our discussion, as he says that alleviating racial stratification "is a more acute interest than diversity" (Huq 2019, 1131); our arguments for a social impact approach rest on justifications based directly with alleviating racial stratification, and not with the "diversity" goals sometimes associated with affirmative action policies.

Even in business ethics, where values are considered, one dominant view is that corporations should be managed in the best interests of shareholders (Moriarty 2021). If firms should be managed in the best interests of shareholders, then the use of an algorithm that produces disparate impacts for marginalized groups would perhaps be seen as defensible in cases where the company derives a benefit from its use. For instance, if credit score is being used as a proxy for income, and Black people in the US end up paying higher premiums than their white counterparts because of correlations, we've shown this could be actuarially fair, and from a shareholder's perspective may be defensible: poor people tend to file more claims and consequently impose a higher expected cost. Mitigating disparate impacts for marginalized communities could be antithetical to shareholders' interests.

In response, notice first that many companies have value statements that indicate their endorsement of antiracist and equity-related values. In the wake of Black Lives Matter protests and other forms of antiracist activism, corporations have shared messages on social media condemning racism and announcing their donations to advocacy groups like the NAACP Legal Defense and Educational Fund and the Equal Justice Initiative. Some corporations have pledged to take actions to divest from the harmful systems of American policing. For instance, IBM and Microsoft said they would no longer sell facial recognition software to law enforcement until further notice or until there are laws regulating it (Bensinger 2020). For companies expressing antiracist values, a social-impact framework could be relevant and useful to guiding action.

But more broadly, as Jeffrey Moriarty (2021) points out, shareholder primacy and other theories of business ethics (such as stakeholder theory) should not be interpreted as views about the ultimate ends of decision-making: managers should also make decisions that are consistent with the requirements of morality. Working out those requirements is part of what business ethics is all about. Notably, many standard business ethics textbooks do not include discussions of topics related to racial equity, such as "retail redlining"—a concept that predates algorithms and refers to "inequitable distribution of retail resources across racially distinct areas" (Kwate et al. 2013). Plausibly, antiracism falls within the purview of moral requirements and constraints; these reflections on the social impacts of algorithms thus suggest that racial equity and disparate impact should be a topic of increased attention in business ethics research and teaching.

It's worth noting in this context that a potentially constructive aspect of social-impact framing—and of the structural theories it emerges from—is that they allow for shifting our attention away from individual blame and more toward collective responsibility. Frameworks focused on intention tend to encourage an emphasis on individual blame and the question of when it is appropriate to hold particular persons responsible for discriminatory or oppressive effects. Holding people responsible can be important, but identifying individuals to blame can be difficult, since it requires us to consider the extent to which we are individually morally blameworthy for perpetuating discrimination—and, in the case of unintentional proxy discrimination, possibly unknowingly. Even defining an individual's role in a large bureaucratic structure can be difficult. But when we shift our focus away from intent to something more systemic, we are better able to explain directly why algorithms can be discriminatory, and why there is an obligation to address this discrimination, without needing a judgment about who is specifically to blame. One possibility in this direction is that the responsibility associated with adopting a social-impact framework could be understood as a collective responsibility. Forward-looking collective responsibility can be described as a responsibility for making sure that a particular desirable state of affairs comes into existence (see, e.g., French and Wettstein 2014). In this context, forward-looking collective responsibility concerns what collective agents should be doing to make conditions better and could focus our attention on the shared burden to eliminate racism.

Conclusion

We've argued that from the perspective of structural theories of racism, what makes a given use of an algorithm discriminatory is best understood in terms not of intentions or fairness-as-parity but rather with respect to social impact. In particular, we should consider how an algorithm's use reinforces or ameliorates existing social stratification. In contrast to some existing definitions of algorithmic bias, this characterization is asymmetrical with respect to different groups. We've suggested that this framing points toward interesting avenues for further investigation in business ethics related to racial equity more generally and that it allows for a potentially constructive view of accountability based on collective responsibility.

References

Ajunwa, Ifeoma. 2020. "The Paradox of Automation as Anti-Bias Intervention." *Cardozo Law Review* 41, no. 5 (June): 1677–742. https://cardozolawreview.com/the-paradox-of-automation-as-anti-bias-intervention/.

Barocas, Solon, and Andrew D. Selbst. 2016. "Big Data's Disparate Impact." *California Law Review* 104, no. 3 (June): 671–732. https://www.californialawreview.org/print/2-big-data/.

Bensinger, Greg. 2020. "Corporate America Says Black Lives Matter. It Needs to Hold Up a Mirror." *New York Times*, June 15, 2020. https://www.nytimes.com/2020/06/15/opinion/black-lives-matter-corporate-pledges.html.

Bent, Jason R. 2020. "Is Algorithmic Affirmative Action Legal." *Georgetown Law Journal* 108, no. 4 (April): 803–53. https://www.law.georgetown.edu/george town-law-journal/in-print/volume-108/volume-108-issue-4-april-2020/is-algorithmic-affirmative-action-legal/.

Birhane, Abeba, Elayne Ruane, Thomas Laurent, Matthew S. Brown, Johnathan Flowers, Anthony Ventresque, and Christopher L. Dancy. 2022. "The Forgotten Margins of AI Ethics." In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency*, 948–58. New York: Association for Computing Machinery. https://doi.org/10.1145/3531146.3533157.

Chander, Anupam. 2017. "The Racist Algorithm?" *Michigan Law Review* 115 (6): 1023–45. https://doi.org/10.36644/mlr.115.6.racist.

Datta, Amit, Michael Carl Tschantz, and Anupam Datta. 2015. "Automated Experiments on Ad Privacy Settings." In *Proceedings on Privacy Enhancing Technologies* 2015 (1): 92–112. https://doi.org/10.1515/popets-2015-0007.

Fazelpour, Sina, and Zachary C. Lipton. 2020. "Algorithmic Fairness from a Non-ideal Perspective." In *AIES '20: Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society*, 57–63. New York: Association for Computing Machinery. https://doi.org/10.1145/3375627.3375828.

French, Peter A., and Howard K. Wettstein, eds. 2014. *Forward-Looking Collective Responsibility*. Boston: Wiley.

Frye, Marilyn. "Sexism." 1983. In *The Politics of Reality: Essays in Feminist Theory*, 17–40. Trumansburg, NY: Crossing Press.

Gandy, Oscar H., Jr. 2016. *Coming to Terms with Chance: Engaging Rational Discrimination and Cumulative Disadvantage*. New York: Routledge.

Hacker, Philipp. 2018. "From Algorithmic Discrimination to Algorithmic Fairness." RAILS: Robotics & AI Law Society blog, October 1, 2018. https://ai-laws.org/en/2018/10/from-algorithmic-discrimination-to-algorithmic-fairness-dr-philipp-hacker-ll-m/.

Heath, Joseph. 2007. "Reasonable Restrictions on Underwriting." In *Insurance Ethics for a More Ethical World*, edited by Patrick Flanagan, Patrick Primeaux, and William Ferguson, 127–59. Bingley, UK: Emerald Group.

Hellman, Deborah. 2020. "Measuring Algorithmic Fairness." *Virginia Law Review* 106, no. 4 (June): 811–66. https://www.virginialawreview.org/articles/measuring-algorithmic-fairness/.

Huq, Aziz Z. 2019. "Racial Equity in Algorithmic Criminal Justice." *Duke Law Journal* 68, no. 6 (March): 1043–134. https://scholarship.law.duke.edu/dlj/vol68/iss6/1/.

Jha, Saurabh. 2012. "Punishing the Lemon: The Ethics of Actuarial Fairness." *Journal of the American College of Radiology* 9, no. 12 (December): 887–93. https://doi.org/10.1016/j.jacr.2012.09.012.

Johnson, Gabbrielle M. 2021. "Algorithmic Bias: on the Implicit Biases of Social Technology." *Synthese* 198, no. 10 (October): 9941–61. https://doi.org/10.1007/s11229-020-02696-y.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Cass R. Sunstein. 2018. "Discrimination in the Age of Algorithms." *Journal of Legal Analysis* 10:113–74. https://doi.org/10.1093/jla/laz001.

Kwate, Naa Oyo A., Ji Meng Loh, Kellee White, and Nelson Saldana. 2013. "Retail Redlining in New York City: Racialized Access to Day-to-Day Retail Resources." *Journal of Urban Health* 90, no. 4 (August): 632–52. https://doi.org/10.1007/s11524-012-9725-3.

Landes, Xavier. 2015. "How Fair is Actuarial Fairness?" *Journal of Business Ethics* 128, no. 3 (May): 519–33. https://doi.org/10.1007/s10551-014-2120-0.

Lebron, Christopher J. 2013. *The Color of Our Shame: Race and Justice in Our Time.* New York: Oxford University Press.

Meyers, Gert, and Ine Van Hoyweghen, I. 2018. "Enacting Actuarial Fairness in Insurance: From Fair Discrimination to Behaviour-Based Fairness." *Science as Culture* 27 (4): 413–38. https://doi.org/10.1080/09505431.2017.1398223.

Mills, Charles W. 2003. "White Supremacy as Sociopolitical System: A Philosophical Perspective." In *Whiteout: The Continuing Significance of Racism,* edited by Ashely "Woody" Doane and Eduardo Bonilla-Silva, 35–48. New York: Routledge.

———. 2005. "'Ideal Theory' as Ideology." *Hypatia* 20, no. 3 (Summer): 165–83. https://doi.org/10.1111/j.1527-2001.2005.tb00493.x.

Moriarty, Jeffrey. 2021. "Business Ethics." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2021 edition. Published November 17, 2016; substantive revision June 8, 2021. https://plato.stanford.edu/archives/fall2021/entries/ethics-business/.

Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: NYU Press.

O'Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown.

Pessach, Dana, and Erez Shmueli. 2020. "Algorithmic Fairness." arXiv preprint, arXiv:2001.09784 [cs.CY]. https://doi.org/10.48550/arXiv.2001.09784.

Prince, Anya E. R., and Daniel Schwarcz. 2020. "Proxy Discrimination in the Age of Artificial Intelligence and Big Data." *Iowa Law Review* 105, no. 3 (March): 1257–318. https://ilr.law.uiowa.edu/print/volume-105-issue-3/proxy-discrimination-in-the-age-of-artificial-intelligence-and-big-data.

Raghavan, Manish, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. "Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices." In *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 469–81. New York: Association for Computing Machinery. https://doi.org/10.1145/3351095.3372828.

Young, Iris Marion. 2011. "Five Faces of Oppression." In *Justice and the Politics of Difference*, ch. 2. Princeton, NJ: Princeton University Press.

ALYSHA KASSAM received her PhD from the Logic and Philosophy of Science program at the University of California, Irvine in 2021. Her research focuses on the role that social and political values play in scientific research, especially in policy-relevant areas. More recently, she has been studying the various ways nonepistemic values are encoded in complex and oftentimes opaque mathematical models across disciplines. She currently teaches courses at California State University Fullerton and California State University Long Beach.

PATRICIA MARINO is professor of philosophy at the University of Waterloo in Canada, where she works in ethics, epistemology, philosophy of economics, and philosophy of sex and love. She is the author of *Moral Reasoning in a Pluralistic World* (McGill-Queens University Press, 2015) and *The Philosophy of Sex and Love: An Opinionated Introduction* (Routledge, 2019) as well as articles on economic methodology, law and economics, bioethics, moral dilemmas, sexual objectification, and other topics. For more information, visit patriciamarino.org.