Where fairness fails:
Data, algorithms, and the limits of antidiscrimination discourse

Anna Lauren Hoffmann
Assistant Professor
The Information School
University of Washington
alho@uw.edu

Abstract: Problems of bias and fairness are central to data justice, as they speak directly to the threat that "big data" and algorithmic decision-making may worsen already existing injustices. In the United States, grappling with these problems has found clearest expression through liberal discourses of rights, due process, and antidiscrimination. Work in this area, however, has tended to overlook certain established limits of antidiscrimination discourses for bringing about the change demanded by social justice. In this paper, I engage three of these limits: 1) an overemphasis on discreet "bad actors", 2) single-axis thinking that centers disadvantage, and 3) an inordinate focus on a limited set of goods. I show that, in mirroring some of antidiscrimination discourse's most problematic tendencies, efforts to achieve fairness and combat algorithmic discrimination fail to address the very hierarchical logic that produces advantaged and disadvantaged subjects in the first place. Finally, I conclude by sketching three paths for future work to better account for the structural conditions against which we come to understand problems of data and unjust discrimination in the first place.

Keywords: big data, algorithms, antidiscrimination, social justice, intersectionality

Introduction

Bias and fairness are central themes in the emerging domain of data justice. As matters of concern, they speak to the threat that "big data" and algorithmic decision-making, when applied to particular sorts of problems, risk worsening already unjust distributions of important liberal goods like rights, opportunities, and wealth. As Mimi Onuoha (2018) describes it, this threat is worrisome not only because it may create new inequalities, "but because it has the power to cloak and amplify existing ones" (n.p.). In the United States, grappling with these difficult problems has found its clearest expression through discourses of rights, due process, and antidiscrimination. For concerned experts in academia, industry, and government alike, these tools (and their attendant histories and literatures) appear to offer a potent counter to systems that, if left unchecked, would happily sort and segregate and optimize without regard to histories of, for example, racial or gendered discrimination. At the same time, however, an uncritical faith in these tools risks reproducing well-documented issues in how ideals of antidiscrimination have been interpreted and applied, especially in legal and political contexts. As both critical legal scholars and critics of liberal political frameworks have pointed out, historical application of these ideals has in many ways hindered antidiscrimination doctrine's effectiveness for bringing about the transformative and lasting structural change that social justice demands.

In this paper, I engage three critiques of antidiscrimination discourse relevant to discussions of data, bias, and fairness today, especially those that have made visible antidiscrimination's shortcomings in addressing injustice relative to race, ability, and liberal goods like rights, opportunities, and material resources. The paper moves three parts. First, I provide a snapshot of data and discrimination concerns, past and present. Second, I sketch and summarize three limits of antidiscrimination discourses both in the law and in liberal political frameworks generally. In each sub-section, I follow the introduction of each critique with a discussion of parallel limits in attempts to address problems of unfairness and bias and data-based discrimination, especially those centered on mitigating the prejudices of imperfect humans behind the machine or efforts to develop computational, policy, or other technical solutions to problems of bias and unfairness. I show that, in mirroring some of antidiscrimination discourse's most problematic tendencies, efforts to achieve fairness and combat algorithmic discrimination fail to address the very hierarchical logic that produces advantaged and disadvantaged subjects in the first place. Instead, these efforts have tended to admit, but place beyond the scope of analysis important structural and social concerns relevant to the realization of data justice. Finally, I conclude by sketching three paths for future work that moves beyond narrowly causal conceptions of algorithmically-mediated systems in order to better account for their role in producing the very (structural and hierarchically-ordered) conditions against which we come to understand problems of data and discrimination in the first place.

Data's discriminatory potential

Concern that technological advances can have unsavory or unjust discriminatory consequences is a persistent theme in the history of computation and data processing. Early efforts to computerize governmental records in the United States during the 1950s and 1960s sparked concerns over the possibility of data collection and processing, absent transparency and proper oversight, fueling the unfair treatment of citizens—concerns that were later extended to private data stores held by credit and medical information reporting agencies (McNamara, Jr., 1973, pp. 71–77). Congressional hearings during this time helped expose unreliable and often dubious collection methods of these agencies, as well as the impact of false or inaccurate data on consumers' social and economic lives. Legislative and other responses sought, among other things, to mitigate the potential for personal data to be amassed and used in discriminatory ways, as with the passage of the Fair Credit Reporting Act and establishment of the Secretary's Advisory Committee on Automated Personal Data Systems. The committee's records indicate that members were particularly concerned with questions of fairness, due process, and transparency (Hoofnagle, 2015).

Subsequent advances in computation and networked information systems renewed concerns over unfair discrimination as a result of systems designed—as Oscar Gandy, Jr. (1995) warned—to surveil, sort, and optimize in the pursuit of social and economic control (p. 37; see also Gandy, Jr., 1993). Similarly, Deborah Johnson and John Mulvey (1993) sketched the emerging problem of bias for computerized decision making, echoing James Moor's (1985) earlier illustration of bias in widely-used airplane reservation software that unfairly advantaged some companies at the expense of others. Using this case as a jumping off point, Batya Friedman and Helen Nissenbaum (1996) advanced their germinal account of computer systems "that systematically and unfairly discriminate" (p. 332). Though this work did not necessarily address discrimination in its particular manifestations and histories, it nonetheless helped connect computational processes with processes of bias and discrimination generally. Later work, like Harcourt's (2006) examination of racial profiling and predictive analytics in policing, would engage more directly with specific intersections of data use and historically-situated discrimination.

Today, work on bias in computing is driven, in part, by the rapid proliferation of devices and online platforms that facilitate the production and capture of massive amounts of social and behavioral data (boyd and Crawford, 2012). In the United States, the Obama White House—presaged by key early scholarship (e.g. Citron, 2007; Sweeney, 2013) and developments overseas (e.g. European Commission, 2012)—confronted the challenges of so-called "big data" through the lens of discrimination and civil rights (Podesta, Pritzker, Moniz, Holdren, & Zientz, 2014). In a series of reports, the administration detailed the "enormous potential for positive impact" and the "unintended discriminatory consequences" of "algorithmic systems and automated processes [that] inform decisions that affect our lives, such as whether or not we qualify for credit or employment opportunities, or which financial, employment and housing advertisements we see" (Muñoz, Smith, & Patil, 2016, pp. 4–5). These efforts helped to further legitimize "data and discrimination" as an area of political, ethical, and economic inquiry (see also Gangadharan, 2014; Data & Civil Rights, 2014).

In the years since, popular texts have probed the negative impacts of big data and algorithmic decision-making in a range of contexts (e.g. O'Neil, 2016). At the same time, legal scholars, journalists, and others have uncovered further issues, showing how algorithmic systems stand to undermine fairness and individual due process rights (e.g. Citron & Pasquale, 2014; Crawford & Schultz, 2014; Angwin, Larson, Mattu, & Kirchner, 2016). In their influential paper, "Big Data's Disparate Impact," Solon Barocas and Andrew Selbst (2016) use the lens of American antidiscrimination law to explore cases of discrimination where algorithmic techniques like data mining go both "wrong" (i.e., where a discriminatory decision issues from incorrect or erroneous data) and "right" (i.e., where overly exact systems make a discriminatory decision by accurately reproducing already existing prejudices) (pp. 729–732). Despite either risk, however, they stress that data mining can and should be "part of a panoply of strategies for combatting discrimination in the workplace and for promoting fair treatment and equality" (Barocas & Selbst, 2016, p. 732).

Among this "panoply of strategies" are efforts to develop computational solutions to problems of bias and unfairness (e.g. Calders & Verwer, 2010; Dwork, Hardt, Pitassi, Reingold, & Zemel, 2012; Kamiran & Calders, 2012; Hajian & Domingo-Ferrer, 2013; Feldman, Friedler, Moeller, Scheidegger, & Venkatasubramanian, 2015; Berk, Heidari, Jabbari, Kearns, & Roth, 2017). Researchers interested in unfair discrimination in machine learning, for example, have sought to advance statistical definitions of fairness operationalizable within computational systems, debating which definitions best serve the ends of antidiscrimination (e.g. Kleinberg, Mullainathan, & Raghavan, 2016; Friedler, Scheidegger, & Venkatasubramanian, 2016;

Kilbertus et al., 2017; Kearns, Neel, Roth, & Wu , 2018). Some definitions, like statistical or demographic parity, hold that fair machine decisions are those that treat the general population roughly similarly to some specified or protected subgroup, even if it means intervening in positive (i.e., affirmative) ways; others, like the equal opportunity constraint, seek parity in the odds of false positive and false negative rates across relevant or protected subgroups. The former echoes discourses of formal equality in outcomes, wherein discrimination is said to be avoided when cases that are alike in some (normatively relevant) respect are treated in like ways. The latter accounts seek some notion of equality of opportunity, arguing that discrimination is avoided when relevant subgroups are not disadvantaged at higher overall rates than others.

Engaging the Limits of Fairness and Antidiscrimination Discourse

Today, it is widely recognized that data and algorithms risk reproducing biases against historically disadvantaged populations in ways that, as Barocas and Selbst (2016) put it, "look a lot like discrimination" (p. 673). In the United States, this risk—and efforts to mitigate it—have in many ways echoed liberal antidiscrimination discourses in the law, which have historically sought to address injustices in the distribution and exercise of important rights, opportunities, and resources in domains like voting, housing, and employment. But the degree to which the law has succeeded in these efforts is an open and ongoing question. In particular, certain well documented tendencies in the way courts have interpreted ideals like fairness and antidiscrimination have arguably hindered its effectiveness. These tendencies point toward (perhaps fatal) limits of antidiscrimination discourse for realizing social justice in any broad or meaningful way—limits that extant work on data and discrimination risk inheriting.

Below, I briefly sketch three of these tendencies—1) an emphasis on discreet "bad actors," 2) single-axis thinking and the centering of disadvantage, and 3) inordinate focus on a limited set of goods—paying particular attention to their sources, motivations, and consequences. I then demonstrate parallel limits in attempts to address problems of unfairness and bias and data-based discrimination, especially those centered on mitigating the prejudices of imperfect humans behind the machine or developing computational solutions to problems of bias and unfairness. It is worth noting one limitation of the following discussion: its focus is exclusively on distinctly U.S.-based discourses of fairness and antidiscrimination. However, my aim here is not a full accounting of these critiques and the bodies of work that have grown up around them; rather, I simply mean to draw out some salient features of these discussions that shed light on similar tendencies in efforts to combat data and discrimination.

From "Bad Actors" to "Bad Algorithms"

The first tendency centers on the law's concern with neutralizing inappropriate conduct on the part of individual perpetrators—what I call the "bad actor" frame. In his pathbreaking article "Legitimizing Racial Discrimination Through Antidiscrimination Law," Alan David Freeman (1978) showed United States Supreme Court doctrine was hindered by limited interpretations of "fault" and "causation" that tilted antidiscrimination law away from its broader aims of social emancipation and towards a kind of narrow, mechanistic reasoning. Instead of addressing pernicious social and systemic injustices, the notion of "fault" had, according to Freeman (1978), morphed into a requirement to identify and near-surgically separate from the sum total of possible sources of discrimination only "those blameworthy individuals…violating an otherwise shared norm" (pp. 1053–1054). Similarly, the idea of "causation" bracketed broad social or systemic issues, instead allowing only claims that could be demonstrated in narrow cause-and-effect terms (Freeman, 1978, p. 1056). Working in tandem, these requirements placed on victims the onerous and often impossible burden of isolating only those conditions mechanically linked to discreet and "blameworthy" perpetrators, "regardless of whether other conditions…would have to be remedied for the outcome of the case to make any difference at all" (Freeman, 1978, p. 1056).

For Freeman (1978), this model generated at least two unsatisfactory outcomes. First, the reliance on a narrow cause-and-effect conception of discrimination stripped discriminatory events of their broader social or cultural context, positing discrimination not as a social phenomenon "but merely as the misguided conduct of particular actors" (Freeman, 1978, p. 1054). Or, as Neil Gotanda (1991) describes it, "despite the fact that personal racial prejudices have social origins, racism is considered an individual and personal trait" (p. 44). As a consequence, important social and contextual issues were left "beyond" the law's reach, including discriminatory conduct that may be unintentional or prejudicial actions that could not be easily traced to a discreet discriminatory effect (Freeman, 1978, p. 1056). Second, the model positions the problem of discrimination not as something that will not be solved until all contributing conditions have been eliminated, but, rather, as a matter of neutralizing inappropriate conduct on the part of individual perpetrators (Freeman, 1978, p. 1053).

Similarly, combatting discrimination in the context of data and algorithmic systems means identifying specific data sources, technical features, or human biases at the root of particular unfair outcomes. Sometimes this means looking at the decisions of specific designers or the demographic composition of engineering or data science teams to identify their social "blindspots" (Snow, 2018). The idea here is that, as one *New York Times* article put it, "software is not free of human influence" because "algorithms are written and maintained by people" (Miller, 2015, n.p.). But just as the search for bad actors places structural issues beyond the law's reach, appealing to the "blindspots" of particular designers or teams ignores the structuring role of technology, instead reducing a system's shortcomings to the biases of its imperfect human designers.

Of course, not all work in this area reduces discrimination entirely to some set of "blameworthy" humans behind the machine. Many discussions make clear that algorithmic discrimination can happen in ways that are unintentional or difficult to account for, for example when upstream social biases are reflected in training data in ways that may be difficult to predict. In these cases, biases are said to "sneak in" (University of Bath, 2017), "whether on purpose or by accident" (Barocas & Selbst, 2016, p. 674), or in ways that only emerge over time (Friedman & Nissenbaum, 1996). As one White House report put it, there is an acute risk of "unintentional perpetuation and promotion of historical biases," especially in cases "where a feedback loop causes bias in inputs or results of the past to replicate itself in the outputs of an algorithmic system" (Muñoz et al., 2016, p. 8). But de-emphasizing intentionality does not automatically move us beyond a discreet source mentality, as it still emphasizes a kind of technical causation that simply replaces "bad actors" with "bad data" or "bad algorithms." It still permits ignorance of the ways humans and technology co-conspire to not just passively reproduce but actively uphold and reproduce discriminatory social structures, especially in the case of negative externalities "learned" by, for example, machine learning systems based on subsequent user interaction (Overdorf, Kulynych, Balsa, Troncoso, & Gürses, 2018). But for most "fair" solutions, sources of discrimination that cannot be traced to discreet bad mechanisms are bracketed, dismissed as someone else's problem or, worse, couched as untouchable facts of history—mere accidents that are "caused," if at all, by biases that "sneak in" to the system.

It is important to note that, despite these limitations, our understanding of structural and other conditions that contribute to discrimination has improved greatly in recent decades, in part a result of what Samuel Bagenstos (2006) calls "the structural turn" in antidiscrimination scholarship (p. 2). For example, social psychological research into the nature and pervasiveness of "unconscious" or implicit bias has inspired legal scholars to better account for unintentional forms of discrimination—that is, prejudicial associations between particular attributes and particular social groups that are not apparent or even consciously available to individuals (e.g. Krieger, 1995). Obviously, the existence of unconscious bias calls into question the "bad actor" frame's dependency on narrow causal links and "blameworthy" perpetrators. In the context of the law, however, Bagenstos (2006) notes that these efforts, rather than liberating antidiscrimination law from its former constraints, are instead further evidence of the law's limits. In a technology context, engineers and technology firms have started to gesture towards unconscious bias as that intangible something else we ought to address. Here, bias is externalized and transformed into something that, as Linda Hamilton Krieger (1995) once put it, "sneak[s] up on" us from the outside (p. 1188), as opposed to something that is variously, but systematically cultivated and maintained. The idea that our biases are somehow apart from us yet can infect our decision-making converts them into something akin to what Freeman (1978) sarcastically called our "ancestral demons"—that is, a possession or invasion for which we are not at fault but which we should nonetheless seek to purge. In this space, unconscious bias training programs pick up where technical fixes leave off: rather than take responsibility for the ways we are daily and actively complicit in reifying culturally-situated violences, we externalize bias and—after a few all-day seminars—count our demons exorcised.

Taking intersectionality seriously

The second limit is a tendency of antidiscrimination discourse towards single-axis thinking centered on disadvantage. This insight is central to Kimberlé Crenshaw's influential notion of "intersectionality" (1989; 1991) and the law's role in producing the very social categories through which we come to understand and adjudicate discrimination. Crenshaw's (1989) foundational work showed that the propensity of courts to focus on one axis of discrimination at a time—for example, race or gender—made vulnerable those whose experiences were not reducible to one axis or another. In her paradigmatic example, Crenshaw found that Black women in employment discrimination cases were unsuccessful, in part, because courts compared their claims against the experiences of similarly situated Black men (for racial discrimination cases) or white women (for gender discrimination)—both groups that, unlike Black women, enjoy systematic advantages along at least one historically-contingent dimension (gender, for Black men; race, for white women).

Despite broader debates about the possibilities and limits of intersectionality as a theoretical lens (e.g. Mutua, 2006; Nash, 2017), Crenshaw's basic insight remains indispensable: Black women are vulnerable to discrimination not merely by virtue of being Black women, but because the law's single-axis thinking explicitly produces vulnerabilities for those who, like Black women, are multiply-oppressed. Accordingly, thinking intersectionally is not—as it is sometimes couched—simply a matter of "stacking" or counting oppressions and arriving at a sum total of disadvantages; nor does it mean adopting a quasi-positivist stance that regards identity categories as static or pre-given (Carastathis, 2016, p. 4). Rather, recalling Crenshaw's often-overlooked metaphor of the basement, intersectionality is concerned with both the production and hierarchical ordering of identity categories, as in the above example where the law's single-axis thinking promotes the interests of some and relegates others to the "basement" of the social hierarchy (Carastathis, 2016). Accordingly, intersectionality does not seek some "flat geography" (McKittrick, 2006) of identity categories, but aims to map the contingencies of particular institutional or social arrangements.

Single-axis thinking also tends to focus on relative disadvantage at the expense of attention to the production of systematic benefits or privileges. Factors like race or sex tend to only become salient when they explicitly disadvantage victims, while the privileging of, for example, whiteness or maleness itself is not made explicit (Crenshaw, 1989, p. 151). As Barbara Flagg (1993) describes with regard to race, there is a "tendency of whites not to think about whiteness, or about norms, behaviors, experiences, or perspectives that are white-specific" (p. 957). Moreover, the tendency to associate disadvantage with particular (pre-given) groups represents, as Vivian May (2015) argues, a kind of "'special case' closed logic" that foregrounds comparison across groups at the expense of attention to relevant intragroup differences (p. 26). Black womanhood, for example, is "neither singular nor monolithic," but the focus on disadvantage works to reproduce homogenizing stereotypes (May, 2015, p. 26). Instead of treating as morally abhorrent those structural processes that unjustly advantage certain groups, the focus on disadvantage forces us into a kind of benevolent—or, worse, patronizing—stance that flattens our understanding of those already relegated to the "basement" of the social hierarchy.

As with the law, work on data and discrimination has been slow to fully absorb the lessons of intersectionality. Single-axis thinking is pervasive in efforts to isolate and identify discriminatory biases in data-intensive systems. As Kearns et al. (2017) point out, for example, work on fairness in machine learning tends to focus only on small sets of pre-defined protected attributes (like race or gender). Against this, they propose methods for identifying various combinations of protected attributes and certifying fairness across "exponentially many" subgroups (p. 1). But this move, while an improvement, still falls short as an "intersectional" approach. Intersectionality is not a matter of randomly combining infinite variables to see what "disadvantages" fall out; rather, it is about mapping the production and contingency of social categories. As Buolamwini and Gebru (2018) have shown in their work on facial recognition systems, problematic distinctions between groups are not limited to pre-existing categories but may be produced through interactions between labels in a system (as with labels for race and gender). Despite claims that both works are "intersectional" in the sense employed by Crenshaw (Mitchell, Potash, & Barocas, 2018), only Buolamwini and Gebru (2018) begin to gesture toward the social contingency of difference with which intersectionality is concerned. However, both fall short in terms of addressing the institutionalization of (and liberation from) social hierarchy. For example, Buolamwini and Gebru's (2018) stated aim of "increasing phenotypic and demographic representation in face datasets and algorithmic evaluation" (p. 12) does not address justice issues that arise from the institutional contexts within which facial recognition is employed—a limitation Buolamwini (2018) and Gebru (AI Now Institute, 2018) have emphasized elsewhere.

Further, computational solutions to problems of fairness almost exclusively focus on disadvantage. Zafar et al. (2017), for example, seek to mitigate problems of misclassification across groups (i.e., making sure that women are not misclassified more often than men) in order to ensure that no one group is "mistreated" or put "at an unfair disadvantage", a result of their assumption that there is a "ground truth" of static and pre-given—rather than contingent and constructed—social categories (p. 1). The seductiveness of disadvantage as a focus is also evident in foundational work. For example, the opening illustration in Friedman and Nissenbaum's (1996) germinal "Bias in Computer Systems" features airline reservation software criticized for systematically benefiting one airline at the expense of others, while the reminder of the article addresses systematic disadvantage. The shift is subtle, but consequential: by centering disadvantage, we fail to question the normative conditions that produce—and promote the qualities or interests of—advantaged subjects.

In other cases, discrimination is recast in painfully neutral terms—i.e., "non-discrimination." For Hardt et al. (2016), for example, "non-discrimination" is conceived as rough parity in false negative and false positive rates across protected groups. But while this may appear to move us beyond mere disadvantage, it does little to address the different real-world consequences false results might have for different groups. A person of relative socioeconomic advantage is more likely to have the time or resources necessary to contest an unfair decision—an imbalance that persists regardless of the fact that differently-situated groups stood an equal chance of being falsely flagged within the system. Instead of grappling with the processes that generate patterns of advantage and disadvantage within and across groups, both disadvantage-focused and "non-discrimination" approaches limit us to solutions that are, at best, reactive and superficial. Or, as Virginia Eubanks (2018) summarizes, "when automated decision-making tools are not built to explicitly dismantle structural inequalities, their increased speed and vast scale intensify them dramatically" (Paragraph 17).

Data-based discrimination beyond distributions

The last limitation centers on the tendency of antidiscrimination discourses to focus on disadvantage relative to a narrow set of goods, namely rights, opportunities, and resources. As critics of distributive justice have long shown, near exclusive focus on these goods cannot account for justice issues related the design of social, economic, and physical institutions that structure decision-making power and shape normative standards of identity and behavior

(Young, 2006). Disabilities scholars, for example, have shown how normative standards of ability shape our world in ways that are biased, as when buildings without wheelchair access impose a normative standard of mobility that excludes many otherwise capable persons. Mere focus on distributions of goods, then, fails to account for the way background features of the world—like the built environment—structure the ability to use goods in the first place. Further, as Shew (2018) points out, disability hinges not only on the distribution of assistive technologies or accommodating designs, but also on their maintenance and the social meanings attached to them (see also Bell, 2010; Garland-Thomson, 2006). Social attitudes, for example, play a significant role in shaping persons' well-being in ways that are relevant to the realization of justice, but addressing them is not wholly reducible to matters of redistribution.

Of course, these critiques would ring hollow if centering rights, opportunities, and material wealth had resulted in appreciable and unequivocal gains for those most vulnerable to discrimination or violence. It is far from clear, however, that this is the case. Contrary to the cherry-picked arguments of liberal humanists like Steven Pinker, the tools of conventional liberalism have not straightforwardly or evenly improved people's life chances, even in affluent countries. As Jonathan Gray (2011) notes, these arguments tend to downplay features of contemporary social life—like the persistence and resilience of racism and xenophobia—which suggest that "outside of some fairly narrowly defined areas of scientific investigation, progress is at best fitful and elusive" (n.p.). Indeed, one of the most obvious arguments against centering rights, opportunities, and resources is that—for those groups that continue to be relegated to Crenshaw's "basement"—doing so has not worked. Liberalism's promises of equal rights and freedom from domination continue—as Charles Mills (2017) describes—to allude many, especially people of color. In the United States, violence against people of color persists (especially in the form of mass incarceration, e.g. Alexander, 2012) and racial wealth gaps continue to grow (Spade, 2015, p. 81). This is not to say that a focus on these goods can never produce positive outcomes for oppressed people, nor that they should not be components of future solutions. Rather, it is only to point out that, on their own, they have not proven absolute forces for progress.

A cursory survey of work on data and discrimination reveals focus areas firmly rooted in the liberal rubric of rights, opportunities, and material resources. They include, but are not limited to: freedom of expression, especially automated moderation of online content (e.g. Gillespie, 2012); criminal justice, especially in policing and sentencing (e.g. Harcourt, 2006; Rosenblat, Randhava, boyd, Gangadharan, & Yu, 2014b); employment, especially hiring and surveillance (e.g. Barocas & Selbst, 2016; Levy, 2015; Ajunwa, 2019); consumer protection, especially privacy and consumer choice (Calo, 2013; Rosenblat et al., 2014a): education, especially admissions and evaluation (Fontaine, 2016); and finance, notably credit risk and market manipulation (Pasquale, 2016). The focus on discrimination only as it relates to particular distributive outcomes is also evident in Friedman and Nissenbaum's (1996) argument that "systematic discrimination does not establish bias unless it is joined with an unfair outcome" (p. 333). But an outsized focus on these goods obscures dimensions of justice not easily reconciled with a rubric of rights, opportunities, and wealth. In particular, emphasis on distributions of these goods fails to appropriately attend to the legitimating, discursive, or dignitary dimensions of data and information in its social and political context (Hoffmann, 2015; Hoffmann, 2017; Dencik, Jansen, & Metcalfe, 2018). Moreover, representational or intimate harms are not easily or intuitively remedied, as made clear by Safiya Noble (2018) and her account of Google search's explicit and degrading results for the query "black girls." Money lost can be replaced and rights violated can be restored, but corporate apologies, subtle tweaks to a system, or even financial compensation ring hollow in the face of attacks on one's dignity.

Finally, conversations that center rights, opportunities, and resources also often implicitly position data and algorithms as primarily instrumental in nature. This implicit assumption, as Nick Seaver (2017) notes, recognizes that data and algorithms both shape and are shaped by cultural context, but it casts them as akin to a rock in a stream: "the rock is not part of the stream, though the stream may jostle and erode it and the rock may produce ripples and eddies in the stream" (p. 4). On this model, addressing discrimination requires a clear distinction between "the algorithm" or "the data" and all other sources of discrimination and connecting "the algorithm" or "the data" to a discreet effect or "ripple in the stream." Doing so is, of course, exceedingly difficult, as algorithms operating in networked information environments are, in Mike Ananny's (2016) words, "moving targets" (p. 108). They incorporate real-time feedback from various sources and optimize over variables that are constantly changing, sometimes in unpredictable ways (Overdorf et al., 2018). Accordingly, algorithmic or automated systems do not only issue decisions, they are also intertwined in the production of social and cultural meaning. As André Brock (2018) argues, cultural artifacts like race are not pre-given variables that are simply plugged in to online platforms or technical systems—rather, they are actively mediated by the system's "computational, network, and semantic qualities" (p. 1025).

In this way, data and algorithms do not merely shape distributive outcomes, but they are also intimately bound up in the production of particular kinds of meaning, reinforcing certain discursive frames over others (e.g. Willson, 2017; Sweeney, 2016; Bivens & Hoque, 2018). For example, as Julie Cohen (2018) has demonstrated, extant framings of personal data as both available and potentially valuable sets up a particular kind of relationship between people and those entities that seek to capitalize on people's data, namely one "supports the reorganization of sociotechnical activity in ways directed toward extraction and appropriation" (p. 214). But this normative frame is not an additional "good" to set alongside rights, opportunities, and wealth—rather, it informs the very backdrop against which we understand those goods and ideals of their fair distribution.

Conclusion

In his 2010 article "Engaging Rational Discrimination," Oscar Gandy, Jr. lamented that, after nearly half a century of work on privacy and surveillance studies, he is resigned to the fact that neither "are adequate to the task of managing a system whose purpose is discrimination" (p. 31). Later, he points to "a substantial literature and a well-documented history of civil rights and anti-discrimination legislation" as other sources from which to draw inspiration for combatting the biases and distortions endemic to data-intensive systems explicitly designed to sort, segregate, and optimize (Gandy, Jr., 2010, p. 32). These sources, he suggests, represent an "effort to correct the bias and distortion that prejudice, disregard, animus, and own-group favoritism by humans often introduce into the calculus of social choice" (Gandy, Jr., 2010, p. 32).

But just as attending to privacy is not a panacea for addressing data-based harms, the literature on antidiscrimination law makes clear that certain prominent interpretations of antidiscrimination carry their own limitations—limitations that also plague work on problems of data and discrimination. For Alan David Freeman, an overemphasis on "blameworthy" perpetrators worked to tilt the law towards a narrow conception of discrimination contingent on the identification and isolation of discreet perpetrators "mechanically linked" to discreet discriminatory outcomes. Similarly, 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. Further, Kimberlé Crenshaw's work shows how single-axis thinking in the law actively produces vulnerabilities for certain groups—specifically Black women—while also overfocusing on disadvantage, thus obscuring the production of systematic advantage. Efforts to design and audit algorithmic systems in the name of "fairness" have been hindered by a similar one-dimensional, disadvantage-centered focus. Last, critics of distributive conceptions of justice further show that exclusively attending to goods like rights, opportunities, and material resources—while important—are not sufficient for dismantling or upending these hierarchies. As with Freeman's observation that the law placed structural concerns "beyond" consideration, the outsize focus on a limited set of goods downplays the role of social attitudes and background norm-setting in shaping not only people's well-being, but our very ability to conceive and pursue particular visions of justice.

By noting these limits, I do not mean to suggest that they are absolute or impossible to overcome. Indeed, in suggesting and probing data and discrimination's silences and failures, I am not offering some fatal diagnosis, but rather a generative marking off point. However, the problems sketched in the foregoing sections resist any easy solutions—there is no easy "fix" to be applied at the level of code or in the collection and labeling of training data. Rather, these problems demand sustained and iterative critical attention not only to system failures, but also to the kinds of worlds being built—both explicitly and implicitly—by and through the design, development, and implementation of data-intensive, algorithmically-mediated systems. In that sense, the natural "next steps" that follow from this paper are threefold. First, overcoming current limits requires increased attention to the broader institutional, contextual, and social orders instantiated by algorithmically mediated systems and their logics of reduction and optimization (see: Cinnamon, 2017; Dencik et al., 2018; Overdorf, et al., 2018). Despite widespread recognition that data and algorithms can—by inheritance or on accident—reproduce biases that further disadvantage groups already understood as marginalized, work too firmly rooted in liberal antidiscrimination discourse will likely fail to grapple with the logic of advantage/disadvantage itself. Echoing Freeman, they will force these broader social and political problems "beyond" or outside of active consideration. Second, we need to broaden our scope to better account for the (re)production of the full range of social hierarchy—that is, we must move beyond analyses that center and scrutinize conditions of relative disadvantage to also account for the normalization and production of systematic advantage. Rather than simply a matter of piling up or counting oppressions and their attendant disadvantages, heeding the lesson of intersectionality means paying careful attention to the ways structural processes produce and maintain social hierarchies, normatively promoting the qualities and interests of some while relegating others to society's "basement" (see: Noble, 2016; Costanza-Chock, 2018).

Finally, we must actively resist the view that data and algorithms merely inform, support, or issue decisions that impact distributions of particular goods. In particular, we must confront directly the role data and algorithms play in actively mediating and normalizing the discourses and social conditions against which decisions about distributions can be made in the first place. Kate Crawford and Vladan Joler (2018) put the point in starker terms: "Many of the assumptions about human life made by machine learning systems are narrow, normative and laden with error. Yet they are inscribing and building those assumptions into a new world, and will increasingly play a role in how opportunities, wealth, and knowledge are distributed" (n.p.). By way of contribution to this broader discussion, I have endeavored to show how an uncritical mirroring of the limits of liberal antidiscrimination discourses risks undermining efforts to move beyond talk of "bad data" and "bad algorithms" and towards an intersectional commitment to upending the processes by which institutions, norms, systems generate unjust social hierarchies. As Sasha Costanza-Chock (2018) worries, such a failure will, "through the mundane and relentless repetition of reduction in a thousand daily interactions," simply "produce systems that erase those of us on the margins, whether intentionally or not" (n.p.). At best, we will end up with little more than a set of reactionary technical solutions that ultimately fail to displace the underlying logic that produce unjust hierarchies of better and worse off subjects in the first place.

Acknowledgments

Many thanks to Megan Finn, Sorelle Friedler, Seda Gürses, Alex Hanna, Luke Stark, Daniel Susser, Miriam Sweeney, Nic Weber, and two anonymous reviewers for comments and feedback that informed this work. In addition, this paper would not have been possible without key conversations during the 2017 Lorentz Center workshop on Intersectionality and Algorithmic Discrimination. I am grateful to all participants - and, in particular, Philomena Essed - for their graciousness and insight.

References

AI Now Institute. (2018, November 25). AI Now 2018 Symposium [Video recording]. Retrieved from https://www.youtube.com/watch?v=NmdAtfcmTNg&feature=youtu.be&t=2219

Ajunwa, I. (2019). Algorithms at work: Productivity monitoring platforms and wearable technology as the new data-centric research agenda for employment and labor law. *St. Louis University Law Journal*, *63*(47), 1–47.

Alexander, M. (2012). *The new Jim Crow: Mass incarceration in the age of colorblindness*. New York, NY: The New Press.

Ananny, M. (2016). Toward an Ethics of Algorithms: Convening, Observation, Probability, and Timeliness. *Science, Technology, & Human Values*, *41*(1), 93–117.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine Bias. *ProPublica*. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Bagenstos, S. R. (2006). The structural turn and the limits of antidiscrimination law. *California Law Review*, *94*(1), 1–47.

Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. *California Law Review*, *104*, 671–732.

Bell, C. (2010). Is disability studies actually white disability studies? In L. J. Davis (Ed.), *The Disability Studies Reader* (3rd ed., pp. 266–273). New York, NY: Routledge.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2017). *Fairness in criminal justice risk assessments: The state of the art.* ArXiv:1703.09207 [Stat]. Retrieved from http://arxiv.org/abs/1703.09207

Bivens, R., & Hoque, A. S. (2018). Programming sex, gender, and sexuality: Infrastructural failures in the "feminist" dating app Bumble. *Canadian Journal of Communication*, *43*(3), 441–459.

boyd, d., & Crawford, K. (2012). Critical questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662–679.

Brock, A. (2018). Critical technocultural discourse analysis. *New Media & Society*, *20*(3), 1012–1030.

Buolamwini, J. (2018, June 22). When the robot doesn't see dark skin. *The New York Times*. Retrieved from https://www.nytimes.com/2018/06/21/opinion/facial-analysis-technology-bias.html

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, *81*, 77–91.

Calders, T., & Verwer, S. (2010). Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, *21*(2), 277–292.

Calo, R. (2013). Consumer subject review boards: A thought experiment. *Stanford Law Review*, *66*, 97–102.

Carastathis, A. (2016). *Intersectionality: Origins, contestations, horizons*. Lincoln, NE: University of Nebraska Press.

Citron, D. K. (2007). Technological due process. *Washington University Law Review*, *85*, 1249–1314.

Citron, D. K., & Pasquale, F. (2014). The scored society: Due process for automated predictions. *Washington Law Review*, *89*, 1–34.

Cohen, J. E. (2018). The biopolitical public domain: The legal construction of the surveillance economy. *Philosophy & Technology*, *31*(2), 213–233.

Costanza-Chock, S. (2018). Design justice, A.I., and escape from the matrix of domination. *Journal of Design and Science*, n.p.

Crawford, K., & Joler, V. (2018). Anatomy of an AI System: The Amazon Echo As An Anatomical Map of Human Labor, Data and Planetary Resources [Diagram]. AI Now Institute and Share Lab. Retrieved from http://anatomyof.ai

Crawford, K., & Schultz, J. (2014). Big data and due process: Toward a framework to redress predictive privacy harms. *Boston College Law Review*, *55*, 93–128.

Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A Black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum*, *1*(8).

Crenshaw, K. W. (1988). Race, reform, and retrenchment: Transformation and legitimation in antidiscrimination law. *Harvard Law Review*, *101*(7), 1331–1387.

Data & civil rights: Why "big data" is a civil rights issue. (2014, October 30). Retrieved from http://www.datacivilrights.org/2014/

Dencik, L., Jansen, F., & Metcalfe, P. (2018, August 30). A conceptual framework for approaching social justice in an age of datafication. DATAJUSTICE project. Retrieved from https://datajusticeproject.net/2018/08/30/a-conceptual-framework-for-approaching-social-justice-in-an-age-of-datafication/

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (pp. 214–226). New York, NY: ACM.

Eubanks, V. (2018, January). The digital poorhouse. *Harper's Magazine*. Retrieved from https://harpers.org/archive/2018/01/the-digital-poorhouse/

European Commission. (2012, January 25). Commission proposes a comprehensive reform of data protection rules to increase users' control of their data and to cut costs for businesses. Retrieved from http://europa.eu/rapid/press-release_IP-12-46_en.htm

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *KDD 2015. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 259–268). Sydney, NSW, Australia: ACM Press.

Flagg, B. J. (1993). "Was blind, but now I see": White Race consciousness and the requirement of discriminatory intent. *Michigan Law Review*, *91*(5), 953–1017.

Fontaine, C. (2016, August 8). The myth of accountability: How data (mis)use is reinforcing the problems of public education. *Data & Society*. Retrieved from https://datasociety.net/output/the-myth-of-accountability-how-data-misuse-is-reinforcing-the-problems-of-public-education/

Freeman, A. D. (1978). Legitimizing racial discrimination through antidiscrimination law: A critical review of Supreme Court doctrine. *Minnesota Law Review*, *62*, 1049–1120.

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2016). *On the (im)possibility of fairness*. ArXiv:1609.07236 [Cs, Stat]. Retrieved from http://arxiv.org/abs/1609.07236

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems (TOIS)*, *14*(3), 330–347.

Gandy, O. H., Jr. (1993). *The panoptic sort: A political economy of personal information*. Critical Studies in Communication and in the Cultural Industries. Boulder, CO: Westview Press, Inc.

Gandy, O. H., Jr. (1995). It's discrimination, stupid. In J. Brook & I. Boal (Eds.), *Resisting the virtual life: The culture and politics of information* (pp. 35–47). San Francisco, CA: City Lights.

Gangadharan, S. P. (2014). Data-based discrimination. In S. P. Gangadharan (Ed.), *Data and discrimination: Collected essays* (pp. 2–4). Washington, D.C.: Open Technology Institute.

Garland-Thomson, R. (2006). Ways of staring. *Journal of Visual Culture*, *5*(2), 173–192.

Gillespie, T. (2017). Algorithmically recognizable: Santorum's Google problem, and Google's Santorum problem. *Information, Communication & Society*, *20*(1), 63–80.

Gotanda, N. (1991). A critique of “our Constitution is color-blind.” *Stanford Law Review*, *44*(1), 1–68.

Gray, J. (2011, September 21). Delusions of peace. *Prospect Magazine*. Retrieved from https://www.prospectmagazine.co.uk/magazine/john-gray-steven-pinker-violence-review

Hajian, S., & Domingo-Ferrer, J. (2013). A methodology for direct and indirect discrimination prevention in data mining. *IEEE Transactions on Knowledge and Data Engineering*, *25*(7), 1445–1459.

Harcourt, B. E. (2006). *Against prediction: Profiling, policing, and punishing in an actuarial age* (Reprint ed.). Chicago: University of Chicago Press.

Hardt, M., Price, E., & Srebro, N. (2016, December). *Equality of opportunity in supervised learning*. Paper presented at the 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

Hoffmann, A. L. (2015). Google books, libraries, and self-respect: Information justice beyond distributions. *The Library Quarterly*, *86*(1), 76–92.

Hoffmann, A. L. (2017). Beyond distributions and primary goods: Assessing applications of Rawls in information science and technology literature since 1990. *Journal of the Association for Information Science and Technology*, *68*(7), 1601–1618.

Hoofnagle, C. J. (2015, April 11). The origin of fair information practices [Essay]. Retrieved September 6, 2018, from https://www.law.berkeley.edu

Johnson, D. G., & Mulvey, J. M. (1993). *Computer decisions: Ethical issues of responsibility and bias* (Report No. SOR-93-11). Statistics and Operations Research Series. Princeton, NJ: Princeton University.

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

Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2017). *Preventing fairness gerrymandering: Auditing and learning for subgroup fairness*. ArXiv:1711.05144 [Cs.LG]. Retrieved from http://arxiv.org/abs/1711.05144

Kilbertus, N., Carulla, M. R., Parascandolo, G., Hardt, M., Janzing, D., & Schölkopf, B. (2017, December). *Avoiding discrimination through causal reasoning*. Paper presented at the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, California.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. ArXiv:1609.05807 [cs.LG]. Retrieved from http://arxiv.org/abs/1609.05807

Krieger, L. H. (1995). The content of our categories: A cognitive bias approach to discrimination and equal employment opportunity. *Stanford Law Review*, *47*(6), 1161–1248.

Levy, K. E. C. (2015). The contexts of control: Information, power, and truck-driving work. *The Information Society*, *31*(2), 160–174.

May, V. M. (2015). *Pursuing intersectionality, unsettling dominant imaginaries*. New York: Routledge.

McKittrick, K. (2006). *Demonic grounds: Black women and the cartographies of struggle*. Minneapolis, MN: University of Minnesota Press.

McNamara, R. M., Jr. (1973). The Fair Credit Reporting Act: A legislative overview. *Journal of Public Law*, *22*(1), 67–102.

Miller, C. C. (2015, July 9). When algorithms discriminate. *The New York Times*. Retrieved from https://www.nytimes.com/2015/07/10/upshot/when-algorithms-discriminate.html

Mills, C. W. (2017). *Black rights/white wrongs: The critique of racial liberalism*. Oxford, UK: Oxford University Press.

Mitchell, S., Potash, E., & Barocas, S. (2018). Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. ArXiv:1811.07867 [stat.AP]. Retrieved from http://arxiv.org/abs/1711.05144

Moor, J. H. (1985). What is computer ethics? *Metaphilosophy*, *16*(4), 266–275.

Muñoz, C., Smith, M., & Patil, D. J. (2016). Big data: A report on algorithmic systems, opportunity, and civil rights. Washington, D.C.: Executive Office of the President.

Mutua, A. D. (2006). The rise, development and future directions of critical race theory and related scholarship. *Denver University Law Review*, *84*, 329–394.

Nash, J. C. (2017). Intersectionality and its discontents. *American Quarterly*, *69*(1), 117–129.

Noble, S. U. (2016). A future for intersectional Black feminist technology studies. *The Scholar & Feminist Online*, *13.3-14.1*, n.p.

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York, NY: New York University Press.

O’Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. New York, NY: Crown Publishing Group.

Onuoha, M. (2018, February 7). On algorithmic violence: Attempts at fleshing out the concept of algorithmic violence [Essay]. Retrieved from https://github.com/MimiOnuoha/On-Algorithmic-Violence

Overdorf, R., Kulynych, B., Balsa, E., Troncoso, C., & Gürses, S. (2018, December). *Questioning the assumptions behind fairness solutions*. Paper presented at the 32nd Conference on Neural Information Processing Systems (NIPS 2018), Montreal, Canada.

Pasquale, F. (2016). *The black box society: The secret algorithms that control money and information* (Reprint ed.). Cambridge, MA: Harvard University Press.

Podesta, J., Pritzker, P., Moniz, E. J., Holdren, J., & Zientz, J. (2014). Big Data: Seizing Opportunities, Preserving Values. Washington, D.C.: Executive Office of the President.

Rosenblat, A., Randhava, R., boyd, d., Gangadharan, S. P., & Yu, C. (2014). Data & civil rights: Consumer finance primer [Report]. Data & Society Research Institute. Retrieved from http://www.datacivilrights.org/pubs/2014-1030/Finance.pdf

Rosenblat, A., Wikelius, K., boyd, d., Gangadharan, S. P., & Yu, C. (2014). Data & civil rights: Criminal justice primer [Report]. Data & Society Research Institute. Retrieved from http://www.datacivilrights.org/pubs/2014-1030/CriminalJustice.pdf

Seaver, N. (2017). Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data & Society*, *4*(2), 1–12.

Shew, A. (2017, November 11). Technoableism, cyborg bodies, and Mars [Blog post]. Retrieved from https://techanddisability.com/2017/11/11/technoableism-cyborg-bodies-and-mars/

Snow, J. (2018, February 14). “We’re in a diversity crisis”: Cofounder of Black in AI on what’s poisoning algorithms in our lives. *MIT Technology Review*. Retrieved from https://www.technologyreview.com/s/610192/were-in-a-diversity-crisis-black-in-ais-founder-on-whats-poisoning-the-algorithms-in-our/

Spade, D. (2015). *Normal life: Administrative violence, critical trans politics, and the limits of law* (Revised and expanded ed.). Durham, NC: Duke University Press.

Sweeney, L. (2013). Discrimination in online ad delivery. *Queue*, *11*(3), 10:10–10:29.

Sweeney, M. E. (2016). The Intersectional Interface. In S. U. Noble & B. M. Tynes (Eds.), *The intersectional internet: Race, sex, class, and culture online* (pp. 215–228). Switzerland: Peter Lang International Academic Publishers.

University of Bath (2017, April 13). Biased bots: Human prejudices sneak into AI systems [Press release]. Retrieved from http://www.bath.ac.uk/research/news/2017/04/13/biased-bots-artificial-intelligence/

Willson, M. (2017). Algorithms (and the) everyday. *Information, Communication & Society*, *20*(1), 137–150.

Young, I. M. (2006). Taking the Basic Structure Seriously. *Perspectives on Politics*, *4*(1), 91–97.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., & Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web* (pp. 1171–1180). Perth, Australia: ACM Press.