# How Algorithms Discriminate Based on Data They Lack:

# Challenges, Solutions, and Policy Implications


## Abstract

Organizations often employ data-driven models to inform decisions that can have a significant impact on people's lives (e.g. university admissions, hiring). In order to protect people's privacy and prevent discrimination, these decision-makers may choose to delete or avoid collecting social category data, like sex and race. In this paper, we argue that such censoring can exacerbate discrimination by making biases more difficult to detect. We begin by detailing how computerized decisions can lead to biases in the absence of social category data and in some contexts, may even sustain biases that arise by random chance. We then show how proactively using social category data can help illuminate and combat discriminatory practices, using cases from education and employment that lead to strategies for detecting and preventing discrimination. We conclude that discrimination can occur in any sociotechnical system in which someone decides to use an algorithmic process to inform decision-making, and we offer a set of broader implications for researchers and policymakers.

The idea of being ranked, sorted, and selected based on past records is not new for most people. From university admissions tests and criminal registries to bankruptcy filings and military service records, people expect that records of their actions will have an impact on their futures. However, recent advancements in computational infrastructure and methods are revolutionizing how easily and readily organizations can collect data and perform "data-driven" decisions across institutional contexts. Companies and other institutions can now access years of past records and link a great variety of data sources, sometimes innocuous on their own but not in the aggregate, to inform an increasingly broad range of decisions tied to activities like credit reporting, advertising, and hiring.

In this paper, we contribute to an ongoing public conversation about how data-driven decisions can discriminate by explaining how even unprejudiced computers and decision-makers can generate biased decisions.[1] Although many risks of data collection and storage are well-known, other problems can arise from the refusal to acknowledge or collect certain data. In fact, without social category data, we can ignore or hide, rather than prevent, discrimination, because decisions can be biased even in the absence of social category data. Moreover, in order to check whether such discrimination is taking place, social category data are often needed. When such sensitive information is used responsibly and proactively, ongoing discrimination can be made transparent through data-checking processes that can ultimately improve outcomes for

[1] Romei and Ruggieri provide a thorough technical review of this literature, while a legal overview is provided by Barocas and Selbst, "Big Data's Disparate Impact."

discriminated-against groups.[2] This leaves software and systems engineers, scholars, researchers, and people making decisions for business, education, social services, and other enterprises asking the following questions: When is it appropriate to collect and use sensitive information? When does it cause harm, and when does it prevent harm?

Answering these questions requires understanding the practical and moral sides of a process involving people, data, and computation conceived as a sociotechnical system. Though voices from across sectors and disciplines contribute to this conversation, we bring experience from our own varied social scientific training. Two authors have connections to economics, two to education, one to communication and critical media studies, and all to information science and data studies. Other fields we draw heavily on here are sociology, law, critical race studies, and computer science. We expect few readers to be well-versed in all of these fields, and so we have synthesized this research to present a primer on big data about people, including how computers learn associations from and inform decisions using these data.

[2] Some readers may object that using—and therefore legitimizing and strengthening—social categories harms marginalized groups. Critical race theorists and black feminists identify ethical, analytical, and epistemological problems with using fixed social categories; see Cooper or Gillborn, Warmington, and Demack for a review of these arguments, and see Gandy for arguments in the context of decision-support systems. However, for the purpose of fighting algorithmic discrimination, using categories may be appropriate in some situations, to the extent that they 1) reflect variables available to algorithms, 2) that they are used by powerful organizations, 3) that they reflect groups of people who tend to have similar experiences in some relevant contexts, and 4) that legal redress of discrimination requires their consideration; see Hancock for a more comprehensive argument. We revisit these key critical concepts in Remedy 4 and in the Conclusion.

As a continuing theme, we focus our examples on racial and ethnic bias in United States (U.S.) education and employment. We do so for several reasons. First, racism is complex, harmful, and ongoing. It is pervasive, occurring while people learn skills and begin to work. Witnessing this, readers will better understand how big data can reflect racism and feel the urgency of preventing algorithmic bias. Second, thorough data (often including social categories) on education and employment programs are available through the government, and researchers in these fields often explicitly address bias. Third, readers likely have first-hand experience with education and employment, allowing them more intuition about the scenarios. We hope that this common thread guides readers through the argument, though we also make an effort to show that the discussion relates just as well to many other contexts.

In order to address how bias in computerized decision-making occurs and how to find and fix it, we begin by presenting a primer on how we can think about social categories and big data, and their implications for data privacy. We then provide a detailed explanation of how computerized decisions can be biased, including through *statistical discrimination*. In some situations initially arbitrary biases are reinforced and amplified through feedback effects. To expand on this latter point, we rely on Spence's Nobel Prize-winning economic work known as *signaling theory*. Spence shows that arbitrary biases can arise even when judgments are not explicitly prejudiced and can become self-perpetuating when decision-makers act on these biased judgments.[3] We discuss how such reinforcement can make historical and societal biases "baked into" data-driven, computer-aided decisions.

[3] Spence, "Job Market Signaling."

After revealing the extent of the problem, we show, paradoxically, that to shed light on these biases organizations must collect and carefully use social category data (e.g. about sex and race) – the very data that can be used maliciously to explicitly discriminate. We share four practical cases in which social categories were used proactively to illuminate and combat discriminatory practices. Each case introduces a class of remedies, exploring the benefits and problems tied to each. Our conclusion reviews our interventions proposed as well as challenges further work must address.

## Understanding Social Identifiers in Big Data: A Primer

The former Federal Trade Commission Chair, Edith Ramirez, discusses how "big data" are "assembled, bit-by-bit, from little data,"[4] such as records from service providers and officials. Even when consumers agree to provide a company with their data for one reason, they rarely have control over how it will be used, aggregated, or sold beyond that. Indeed, this is often precisely the specialty of data brokers: they collect and organize data to "create detailed profiles of individuals," and these profiles often necessarily include "highly sensitive information."[5]

Discrimination persists in many people's lives based on a variety of social identifiers. Some jurisdictions legally protect people from differential treatment on the basis of race, ethnicity, religion, national origin, immigration status, sex, gender identity, sexual orientation, disability, and age. Less obvious social categories can also be sensitive, however, including parenthood, military service, involvement with the criminal justice system, political party, and socioeconomic status (SES).

[4] Ramirez, 4.

[5] Ramirez, 7.

In the U.S., many people and institutions discriminate due to prejudice, and still more make decisions tainted by past prejudices. For example, one of the thorniest social problems that America faces is that important life outcomes—from educational attainment and income to incarceration and life expectancy—systematically vary by race. In many cases, scholars have found evidence implicating historical laws and ongoing discriminatory practices.

We can track evidence of racial disparities and discrimination throughout the life course. Upon entering kindergarten, if children are evaluated on how ready they are for school, their scores differ systematically with race, ethnicity, and income.[6] Smith and Harper similarly document widespread disparities in rates of suspension and expulsion from school by race across the southern U.S; they link their findings with prior research showing that black students tended to be disciplined for much more subjective behavior—such as "disrespect"—than white students.[7] When young adults begin to seek jobs, candidates with equivalent job applications or resumes are less likely to be interviewed for a job if their names suggest that they are black rather than white, as audit studies continue to show.[8] During the Great Recession, black homeowners were differentially targeted with predatory mortgage loans and faced greater losses than their white counterparts with similar finances.[9] Many ranking and selection processes in the U.S. today—and their outcomes—are tainted by bias.

[6] Reardon and Portilla; Ready and Wright. These studies are discussed in a scenario later in the paper.

[7] Smith and Harper.

[8] Bertrand and Mullainathan; Agan and Starr; Doleac and Hansen.

[9] Rugh, Albright, and Massey treat the case of predatory lending in Baltimore, Maryland, in detail.

Because of America's past and present discrimination, data about race, ethnicity, immigration status, and gender identity are encoded in many other aspects of a person's life. Where one lives in America remains strongly connected to race[10] and socioeconomic status.[11] Names—both family and personal names—tend to reflect social category membership.[12] Interaction partners—from friends and family to more distant ties—also tend to reflect one's social categories as well as other traits.[13] Increasingly, these bits of information about our lives, suggesting our many social category memberships, are being collected, stored, and used for decision-making across industries.

### *The Nature of Big Data: Full of Correlations*

When traces of people's lives are recorded as "data," and pieced together into "big data," the resulting mesh is densely packed with *correlations*—personal characteristics that tend to show up together. These patterns can exist within a single person's data, revealing themselves as *autocorrelations,* when a single aspect of a person's life is measured repeatedly over time (e.g., last year's income helps predict this year's income). Patterns also exist across people, especially those who interact with one another. These correlations run deep to the extent that we are creatures of habit and that we make choices within the same social systems and shared influences as other people. Databases about people are full of correlations, only some of which

[10] Rugh and Massey.

[11] Reardon and Bischoff.

[12] Gaddis tests perceptions of race suggested by certain names and suggests how to generate representative names.

[13] See McPherson, Smith-Lovin, and Cook for much more about *homophily* in social networks.

meaningfully reflect the individual's actual capacity or needs or merit, and even fewer of which reflect relationships that are causal in nature.

Many computer programs that process data, including so-called "artificial intelligence" and "machine learning" algorithms, learn from patterns. They might be programmed to categorize, score, or make decisions about different people or groups. This means that the densely correlated mesh of personal data has important consequences for how sensitive data are represented and for what such algorithms can do. First, these patterns and correlations make *prediction* possible: if we have enough data about what someone has recently done, and what others around them have done, then we can do a decent job of guessing what they will do in the near future. Second, these patterns and correlations make *imputation* possible: missing data points can be easily inferred by looking at similar people for whom data are available. Third, these patterns and correlations in recorded data also speak to information outside of the dataset: we can closely match data that might be missing through *proxy variables* that are highly indicative of the data that are missing. In the next section, we elaborate on how "big data" allow for better prediction, imputation, and proxy variables with the use of a simple illustration.

### *A Simple Illustration of Big Data*

In order to visualize "big data," we can imagine 1000 American adults, each represented by a piece of graph paper measuring 100 squares by 100 squares. Each column in the paper represents a single question about the person (e.g., how old they are, what state they live in, the year their car was made, their annual income, how long they have worked at their current job, which sports they watch on television, how much they recently spent at a particular online retailer, etc.). The 100 squares in each column represent the possible answers for that particular

question, with the correct answer(s) checked off. Clearly, the pattern of check marks on the graph paper encodes important information about that person's life.

*Data about people are dense with patterns.* If we have the entire stack of these papers, each representing a different person, we may find that some constellations of check marks appear more frequently than others. For example, when people are of similar age and spend similar amounts of money at the same clothing store, they are more likely to have household incomes of a similar level and to consider retiring on similar schedules. By looking through these sheets we might recognize a few different answers that often go together. In fact, it would be very difficult to come up with 100 questions to ask about people where we did NOT see some constellations of answer patterns emerge.

*Patterns across people make prediction possible.* If we wanted to predict who might be ready to find a new job, we could look at the pages for people who recently switched jobs. We could then search for patterns that appear more often among new hires. We might then predict that people with patterns similar to new hires, but who have not as recently switched jobs, may be interested in considering other options.[14]

*Patterns make imputation of missing data possible.* Now imagine that one respondent's record does not have any check marks in the last ten columns. Perhaps there is no record of those particular attributes associated with this person. It would nonetheless be possible to guess values

[14] With intuition or data over time, we might discern other job-hunting patterns: how often does a part-time job lead to a full-time job? Who "job hops" the most? Are there telltale signs that someone will re-enter the workforce, switch industries, or find that their disability prevents them from finding a suitable job?

for that person by looking at the constellations on other sheets that contain comparable answers to the first 90 questions.

*Proxy variables can pinpoint variables that are not represented in the data set.* Now assume that several of the questions were about sensitive data categories, such as racial or ethnic or gender identity. To the extent those categories matter in society, the answers to these questions will be associated with many patterns. These other patterns will provide clues about the answers to the sensitive questions, and vice versa. Many constellations visible in the less sensitive data will reveal insights about more sensitive columns.

Suppose we no longer wanted to have racial, ethnic, or gender identity in the data set. We could erase all check marks from those columns. However, the constellations that are characteristic to certain answers would still be there. Even if an observer did not intend to guess, say, gender identity, they might still detect distinctly gendered patterns.[15]

### *Data Privacy Challenges*

Control over one's sensitive data is valuable, and many people have good reasons for seeking privacy. In particular, divulging sensitive information—even to a trusted entity—may have later repercussions if laws or contracts change. For instance, when a government changes policies about health insurance or immigration, then sensitive information people disclosed under

[15] The example of gender identity suggests the complexity of why attributes and behaviors may be correlated. Some things may be directly linked to performing gender (such as buying a dress) or to its close correlate, sex (such as buying tampons or visiting a gynecologist), while others might correlate for more complex reasons (such as working in a feminized profession or attending yoga classes).

older laws (e.g. pre-existing medical conditions or undocumented immigration status) could prove detrimental.

For example, through the recent U.S. Deferred Action for Childhood Arrivals (DACA) immigration policy, many people lacking legal immigration status registered to legally work and receive renewable, two-year protection from deportation. However, even as immigration policies shift, their data are still held by the government. Legal scholars point out that DACA itself is based on prosecutorial discretion wielded by the executive branch, and its data protections are just as discretionary; thus U.S. President "Trump could rescind existing operation mandates and require USCIS [U.S. Citizenship and Immigration Services] to share this information with the enforcement arms of DHS [Department of Homeland Security]."[16] Cataloging the various "temporary, tenuous, and tentative" versions of U.S. immigration "nonstatus," including deferred action programs, Heeren notes that "nonstatus … offers the government a method of surveillance over the authorized population. One could argue that nonstatus is essentially a registration program."[17] Indeed, some young people who put their home addresses in the registry are experiencing a "'horrific Kafka-like situation' in which they have potentially outed their parents to federal authorities,"[18] as well as risking their own futures now that DACA has been cut short.

Once information is divulged, it can be difficult if not impossible to take it back, and it seems a cruel irony that data solicited under one set of regulations could be used to punish during a subsequent set of regulations. Wariness of such outcomes is part of the impulse behind the

[16] Coutin, Ashar, Chacón, and Lee, p. 958.

[17] Heeren, p. 1132.

[18] Brown, quoting Dr. Marcelo Suárez-Orozco.

Never Again Tech Pledge, drafted and endorsed by many employees of U.S. tech companies. Its signatories vow to "refuse to participate in the creation of [government] databases … to target individuals based on race, religion, or national origin" and to minimize sensitive data collection.[19] After all, can a country that has previously forced its own citizens into internment camps—based on their Japanese origin—truly be trusted not to misuse databases that are labeled by ethnicity, race, nationality, or religion?[20]

When data are "big," unknown data points are more easily filled in through prediction, imputation, and proxies. Consequently, staying private and holding back personal information cannot always prevent this information from being inferred. It can be especially difficult to keep central aspects of one's identity, such as race, gender, or SES, private, as these characteristics are often suggested by many different data traces. Furthermore, withholding information can be a signal in itself. In the economics literature, withholding information is often seen as an attempt not to send a negative signal,[21] and the legal literature has begun to explore the process of privacy "unraveling" as people start explicitly revealing information in order to send a positive

[19] The pledge also encourages them "to scale back existing datasets with unnecessary racial, ethnic, and national origin data." An impetus for this pledge was the worry that the U.S. government could use such data to aid in mass deportations or the internment of immigrants or Muslims. For the full pledge and list of signatories, see http://neveragain.tech/, accessed May 1, 2017.

[20] Anderson and Seltzer trace how the U.S. government's statistical systems separated from its administrative systems, the evolution of "statistical confidentiality," and breaches of this confidentiality, including the (temporarily legal) release of at some Census microdata on individual Japanese-Americans during World War II.

[21] Stiglitz.

signal.[22] Peppet asks, "How long before one's unwillingness to put a monitor in one's car amounts to an admission of bad driving habits, and one's unwillingness to wear a medical monitor leads to insurance penalties for assumed risky behavior?"[23]

There are broader downsides to withholding personal information. Lerman discusses the exclusion of people who are digitally invisible. He points out that billions of people around the world "do not routinely engage in activities that big data and advanced analytics are designed to capture."[24] He goes on to argue that "the nonrandom, systemic omission of people who live on big data's margins, whether due to poverty, geography, or lifestyle,"[25] means that the models of society we create from big data are inevitably biased. Even people who only opt out of certain digital behaviors may not resemble their more-involved peers in important ways. Whether these blind spots affect election polling, social service provision, or how companies understand their potential markets, these systematic omissions can have important impacts. For all these reasons, withholding social category data to keep it private is not necessarily effective or desirable. As we will see in the next section, withholding this information does not stop algorithmic bias – computers produce biased decisions regardless of whether or not they were directed to do so or were given social category information.

## Social Identifiers and Algorithmic Bias

[22] Peppet.

[23] Ibid., 1159.

[24] Lerman, 56.

[25] Ibid., 57.

Algorithms that are designed to find and exploit patterns in big data will pick up on social categories and trace evidence associated with them. However, in many contexts, we as a society believe that membership in a particular social category should not affect how a decision is made, and we are used to situations where we avoid humans' biased judgments by withholding social category information. For instance, many professional U.S. orchestras adopted new audition policies in the 1970s, requiring that candidates play for judges from behind a screen, shifting the focus to be on their performance rather than their gender, race, or familiarity to the judges. Economists Goldin and Rouse found that women who auditioned under both the traditional and "blind" methods were significantly more likely to be advanced when their social category was not known to judges.[26] That is, social norms for fairness altered how a specialized hiring process was conducted, and social scientists later confirmed that blinding auditions lessened bias.

Collecting and accounting for social category data can lessen discrimination, though this idea may seem counterintuitive given the ways societies have historically managed personal information. Here, we make this very case by discussing three main points. These three points, taken together, suggest that when algorithms use "big data" for important decisions, it is futile to exclude social category data.

1. Social identifiers like race and gender are pervasive, such that machine learning algorithms can learn their correlates when trained on past data.

2. The pervasive nature of social identifiers means that such sensitive information is embedded in big datasets, even if it is not intentionally collected or is deleted.

[26] Goldin and Rouse.

3. When an algorithm is fed social category information but is not explicitly designed to avoid discrimination, this can introduce bias into outcomes. This exact situation was modeled over forty years ago and was identified by labor economists as *statistical discrimination.* A special case involving *signaling* shows how feedback effects in a system of data-driven decision-making could create and perpetuate entirely *arbitrary bias*, based on insignificant fluctuations in early data.

Perversely, these three points together mean that algorithms can discriminate on the basis of a social category, intentionally and unintentionally, even when they are not explicitly fed social category data. We elaborate on each of these points in the sections that follow. We distinctly combine these arguments to show how the mechanisms for bias without bigotry fit together with algorithms' prodigious pattern-finding to produce discriminatory results, even when social categories are censored.

### *Computers Learn the Prejudices Connected with Social Identifiers*

Our first point is that social identifiers like race are pervasive, such that machine learning algorithms learn correlates associated with race when trained on past data. A striking example is recent work by Caliskan, Bryson, and Norayanan, who trained an off-the-shelf learning algorithm that associates words that frequently appear together, on a commonly used big dataset of Internet texts. They later replicated the results using a different off-the-shelf algorithm, trained on a different dataset, and tested whether the algorithm held the same implicit word associations that people often do. Indeed, the algorithm replicated common morally-neutral connotations (e.g., flowers are more pleasant than insects, and musical instruments are more pleasant than weapons) and some statistical regularities (e.g., which first names belong to women, men, or both, and which occupations are often held by women or men). In the same way, the algorithm

learned stereotypical biases tied to race and gender. As the authors note, algorithms that are taught broad associations could be prejudiced in making hiring decisions.[27]

Employers might also perform online searches on prospective hires. Sweeney finds that Google serves different ads depending on the name entered in the search box. For example, ads for background checks were more common for names associated with particular races and for males; 60% of the ads offered for black names mentioned "arrest" or "criminal," versus only 48% for white names.[28] She discusses how companies may have requested these ads and how the differences may have been reinforced through systemic feedback.[29]

These studies demonstrate how algorithms can learn negative associations for certain social labels, especially when the data reflect a broad array of inputs. This can also arise within an organization's own data.[30] Barocas and Selbst ask, "How do employers account for the kinds of candidates they have never hired in the past?" This is especially a problem if "past prejudice denied certain classes of candidates the opportunity to demonstrate their talents."[31] As a White House report pointed out, the idea of "hiring for culture fit" could just reproduce past decisions: "Unintentional perpetuation and promotion of historical biases, where a feedback loop causes bias in inputs or results of the past to replicate itself in the outputs of an algorithmic system."[32]

---

[27] Caliskan, Bryson, and Norayanan.

[28] Sweeney also found instances where "arrest" was mentioned for people with no arrest record and where no ad or a neutral ad was offered for people with arrest records.

[29] Sweeney.

[30] Cf. Barocas and Selbst, "Big Data's Disparate Impact," 687.

[31] Barocas and Selbst, "Losing Out on Employment."

[32] Muñoz, Smith, and Patil, 8.

Algorithms learning from big data have plenty of opportunities to associate certain social categories with statistical regularities, stereotypes, and past discrimination.

### *When Sensitive Variables Are Omitted, Computers Still Learn Stereotypes*

Our second point is that algorithms pick out group differences and elements of discrimination in our society. They do so without being able to understand which past outcomes are reliable indicators about a person or group and which are tainted. Importantly, algorithms can even ascertain these social groups when the label itself is not collected or has been deleted. Even algorithms ignorant of identity categories can thus nonetheless act on them, identifying patterns that point to omitted social categories. Indeed, Facebook does not ask users of its social platform about their race or ethnicity. Rather, it guesses a proxy, users' "Multicultural Affinity," based on their site interactions. It then allows advertisers to include or exclude certain groups, which it argues lets companies test different versions of their ads.[33] Advertisers who want to discriminate in hiring or other fields have a tool to do so.

Unfortunately, even if no one explicitly develops a proxy variable, omitted social categories can still drive algorithms. We explain the problem in the context of linear regression

[33] Angwin, Tobin, and Varner. The authors repeated a study done a year earlier, Angwin and Parris, Jr., that asked Facebook to run a housing advertisement that was illegally targeted by race; that first study supposedly led Facebook to make its ad platform comply with the law. However, Angwin, Tobin, and Varner find a year later that all of the targeted ads they submitted were approved, reporting, "The only changes from last year that we could identify in Facebook's ad buying system was that the category called 'Ethnic Affinity' had been renamed 'Multicultural Affinity' and was no longer part of 'Demographics.' It is now designated as part of 'Behaviors.'"

modeling.[34] Leaving out sensitive variables from an analysis forces correlated variables to take on greater significance; these unintentional proxy variables appear to be strong predictors, but they are only so because of their connection with the left-out variable. This result is known as *omitted variable bias*. Pope and Sydnor mathematically explore its consequences,[35] and we later revisit their study as an example of how sensitive information can unmask bias. They point out that when some variables are omitted for being "socially unacceptable for use in predictive models,"[36] their proxies reflect the omitted information, gain heavier use, and might themselves become socially unacceptable. For instance, California car insurance rates must not be set using home locations or credit scores, which too closely reflect the previously-banned factors of race and income.[37] Proxy variables often stand in for omitted categories; the importance of this phenomenon increases as we consider statistical discrimination.

### *Statistical Discrimination and Signaling Theory*

Our third point focuses directly on statistical discrimination, and this section provides a related theoretical discussion about signaling. Decades ago, labor economists began investigating the various mechanisms behind employment discrimination, and one resulting theory is

[34] Although we discuss a parametric model, more complex algorithms suffer from the same fundamental problem unless it is explicitly addressed.

[35] Pope and Sydnor note that they follow in the footsteps of fellow economists Ross and Yinger and Lundberg. From a data mining perspective, Žliobaitė and Custers start with the same problem as Pope and Sydnor but offer different linear regression models for comparison (particularly fitting prediction models separately by subgroup) and extensively consider the legal implications in the European Union.

[36] Pope and Sydnor, 210.

[37] Ibid., 206.

*statistical discrimination*.[38] When employers lack information about ~~an individual job applicant's skills~~an individual job applicant's skill, and there is some cost to hiring the wrong person, they may fill in missing details based on what they know from previous applicants. An extension of this approach, *the signaling model*, shows how feedback effects can sustain unjustified inferences.[39] In this subsection, we introduce these theories and their evidence, explaining how decades-old models can accurately capture the work of cutting edge algorithms.

Hiring is an uncertain process: there is little direct information about how a particular decision will turn out until the new employee actually starts working there. Bluntly, Arrow posits

[38] Arrow "Some Models of Racial Discrimination in the Labor Market"; Phelps. The economic literature on discrimination is too broad to survey here, and it is beyond the scope of this paper to offer a thorough critique of the assumptions made and their plausibility as a proper description of the world, then or now. Arrow ("Some Models") and Phelps both acknowledge that statistical discrimination is just one of the many potential factors in employment discrimination. For example, in 1971 Arrow writes, "Economic explanations for discrimination or other phenomena tend to run in individualistic terms… They tend not to accept as an explanation a statement that employers as a class would gain by discrimination, for they ask what would prevent an individual employer from refusing to discriminate if he prefers and thereby profit. … We must really ask who benefits, and how are the exploitative agreements carried out? In particular, how are the competitive pressures that would undermine them held in check?" (Arrow, "Some Models," 25) Phelps forthrightly acknowledges that it can be difficult to know "whether in fact most discrimination is of the statistical kind studied here. But what if it were? Discrimination is no less damaging to its victims for being statistical. And it is no less important for social policy to counter" (Phelps, 661). In 1998, Arrow revisits racial discrimination, asking, "Can a phenomenon whose manifestations are everywhere in the social world really be understood, even in only one aspect, by the tools of a single discipline?" (Arrow, "What Has Economics to Say about Racial Discrimination?", 91).

[39] Spence, "Job Market Signaling."

that "skin color is a cheap source of information and may therefore be used [to determine a person's likely productivity]," which can replace the undertaking of "a costly operation in information gathering."[40] He speculates as well that "school diplomas are being widely used by employers for exactly that reason, schooling is associated with productivity, and asking for a diploma is an inexpensive operation."[41] As an example, consider a manager who holds no sex-based prejudice, but notices that the firm's past female employees typically stayed with the firm for a shorter amount of time than male employees. Perhaps the manager even identifies an explanation behind the pattern, such as women more often citing family-related reasons for leaving. If the manager uses this group-based evidence to make inferences about future hires' likely tenure, and thus decides to hire fewer women, this is statistical discrimination.[42]

A recent audit study confirms statistical discrimination in hiring, based on online applications to entry-level jobs coded with names that are characteristic of particular races. Agan and Starr find that employers discriminate more on race after laws pass that prohibit them from

[40] Arrow, "Some Models," 21. When he uses the word "may" in this quotation, it is speculative, suggesting what employers might be doing and what might work to some extent in calculations, rather than normative, approving of this race-based generalizing.

[41] Ibid., 21. However, high school diploma requirements, when unrelated to the tasks of the job, were found to be a pretext for racial discrimination in the landmark U.S. Supreme Court case of *Griggs v. Duke Power Co*.

[42] Iverson and Rosenbluth, 4, describe how employees might work longer hours to signal their future productivity, noting that sending such a signal is particularly costly to women because of "extra home duties that society assigns by gender" and mentioning that Spence ("Job Market Signaling") noted exactly this inequality while laying out his theory of signaling.

asking up front about criminal convictions.[43] The authors suggest that employers, relying on perceptions of higher conviction rates of certain races, used race as a proxy to try to avoid applicants with felony records.[44] The employers' low rate of callbacks to black applicants, when they cannot ask about felonies, is so extreme that the authors say it does not seem entirely to be "empirically informed statistical discrimination."[45]

Statistical discrimination models and studies have become extremely relevant with the advent of big data and algorithmic decision-making. When an employer tries to assess a job candidate's future productivity, mostly using records about previous hires, they are faced with a problem different only in scale from what modern decision-making algorithms tackle on a daily basis. The works from the early 1970s even use terms familiar in the Bayesian learning modeling literature:

---

[43] Agan and Starr. Strahilevitz writes about the ethical tradeoff between the privacy of ex-offenders and avoiding discrimination in the labor market, and Agan discusses this in light of recent empirical evidence.

[44] Agan and Starr's study, currently a working paper, uses a careful quantitative methodology, with a triple-difference approach, to audit racial discrimination in two U.S. states, before and after laws about what private employers could ask about took effect. For the audit, researchers created biographical details for fake job applicants to apply to jobs via online forms. The applicants were all men aged 21 or 22, without education beyond high school or a GED. Beyond names (selected to signal being black or white), all applications had similar socioeconomic markers, including similar neighborhoods of residence and high schools attended. The authors explore whether, for people that young, educational attainment and race are indicative of felony convictions; while exact data are difficult to find, their estimates suggest that the true racial gap in felony convictions is far below the gap that would be needed to "rationally" justify the discrimination observed.

[45] Ibid., 28. Doleac and Hansen, with a complementary approach using government data, find consonant results.

> [S]ignals and indices are to be regarded as parameters in shifting conditional probability distributions that define an employer's beliefs. (The shifting of the distributions occurs when new market data are received and conditional probabilities are revised or updated. Hiring in the market is to be regarded as sampling, and revising conditional probabilities as passing from prior to posterior. The whole process is a learning one.)[46]

In short, the processes of learning attributed to a calculating employer in canonical economic models underlie what many algorithms actually implement.

Further, signaling theory shows how feedback effects can sustain even an arbitrary bias. Spence extends this line of work to show that, under some conditions, statistical discrimination can arise *even when there are no underlying differences between various groups*.[47] These results, too, apply precisely to today's decision-making algorithms. Spence's model starts with this same view that hiring is an uncertain investment,[48] emphasizing an individual's incentives to develop *signals* of being more productive. We briefly discuss his simplest model in order to give some intuition for the results. In the model, workers have either high or low productivity at a given job. High productivity allows employers to pay higher hourly wages, and low productivity would probably lead to lower hourly wages. If all workers had the same productivity and everyone

[46] Spence, "Job Market Signaling," 357-8. In the quotation, we append in parentheses the clarification he provided in his footnote 5.

[47] Spence formalizes and extends thoughts from one of his dissertation advisors, Arrow, on how statistical discrimination might allow racial wage gaps—unjustified by individuals' true productivity—to persist (Arrow, "Some Models").

[48] Spence, "Job Market Signaling," 356.

knew it, then wages would be the same for each person. But if high- and low-productivity workers all seem the same at first, then starting wages will tend to be the average of high and low wages, weighted by how many workers of each productivity type are in the whole labor market.

Spence asks what would happen if workers could distinguish themselves. What if there were skill badges were available, which cost low productivity workers twice as much to get as high productivity workers?[49] In pondering the role of such a badge in the marketplace, Spence makes two key assumptions: workers will decide whether to invest in the badge based on its costs and on the wages they would get based on this *signal*, and employers will pay attention to how the badge relates to productivity—the *strength* of the signal—when setting wages.

In one scenario, signaling could reach *equilibrium*—a point where employers' beliefs about what the signal signifies are stable, the beliefs are not "*disconfirmed* by the incoming data and the subsequent experience," and in fact they "are *self-confirming.*"[50] In an equilibrium the wages employers set for workers with and without badges encourage each kind of worker to continue getting badges at a stable rate. Moreover, there is nothing prohibiting there being more than one possible solution, characterized by different badge prices and hiring outcomes. Indeed,

[49] Spence, "Signaling in Retrospect and the Information Structure of Markets," 436. We use "badge" where Spence uses "education," since the "education" process he describes explicitly does not change one's productivity, but instead indicates one was probably already a high productivity type. The badge could cost money, but its "cost" could also reflect the effort, time, or extra resources needed to obtain the badge.

[50] Ibid., 437.

Spence writes that "it is the self-confirming nature of the beliefs that gives rise to the potential presence of *multiple equilibria* in the market."[51]

The idea of multiple equilibria here means that there are multiple potential prices for badges that would encourage all of the high productivity workers to distinguish themselves by getting a badge, but that would remain too costly for low productivity workers to get. If the price of a badge is set somewhere in that range beforehand, then this is a *separating equilibrium* and there is no reason for the price of a badge or wages for either group to change. However, if enough of the population has high productivity (allowing for relatively high wages) and badges cost enough, then instead no one will get a badge and everyone accepts the same wage; this is called a *pooling equilibrium.*[52] There is no telling, in general, whether we will arrive at a separating or pooling equilibrium, and this arbitrariness becomes important when we start to consider people in different social category groups.

Spence takes this next step to show how an arbitrary bias can arise. He specifies that there are two social category groups, each with the *same mixture of high and low productivity*

[51] Ibid., 437. Spence also notes that the employer might have extreme beliefs about productivity that "drive certain groups from the market and into another labor market. … But when it happens, there is no experience forthcoming to the employer to cause him to alter his beliefs" ("Job Market Signaling," 366). This lack of disconfirming evidence can even happen within the same labor market, especially if employers mainly rely on their own proprietary data for their algorithms: as Kim writes, "if the algorithm mistakenly labeled an applicant as 'unqualified,' the employer will not hire her and therefore, will never observe her work performance. As a result, there will be no opportunity to learn of the error and update the model" (Kim, 26).

[52] Spence, "Signaling in Retrospect and the Information Structure of Markets," 438.

*workers.*[53] But even though the situations are symmetric, the situation can evolve differently across the two groups. "One person's signaling strategy or decision affects the market data obtained by the employer,"[54] Spence argues, and from there the employer updates beliefs, wages, and thus applicants' incentives to seek a badge. But the employer might not be sure whether the group identifier—say, gender—matters. A strict empiricist, the employer now conditions beliefs about productivity on both badges and gender, to see if those matter going forward. However, this means that "the external impacts of a man's signaling decision are felt only by other men,"[55] and different beliefs—and thus different wages, incentives, and equilibria—might develop regarding men and women. Spence's elegant model shows that developing and applying decision rules can change applicants' incentives and lead to a stable situation in which people in one social category with high productivity are paid less than people in another social category with high productivity. That is, even without any prejudiced intent and without underlying group differences in productivity, data-driven, rational decision systems can still give rise to inequality.

More recent work on statistical discrimination and signaling tends to focus on the *bounded rationality* of human decision-makers, such as limitations on our memories which keep us from being perfect calculators. Such extensions provide proof of how bias can arise without bigotry in a broader set of contexts. For instance, Varshney and Varshney show that a limited capacity to store fine-grained data can yield racial discrimination even when the groups have the

[53] Spence, "Job Market Signaling," 369. Importantly, because there is no true productivity difference between the groups, statistical discrimination is an inefficient course of action, even separate from an ethical judgment.

[54] Ibid., 370.

[55] Ibid., 370.

same distributions of traits. Their model requires that decision-makers have more experience with one racial group than another, and therefore make finer category distinctions for that group. Indeed, they demonstrate further conditions under which statistical discrimination can develop without any true between-group differences or malice.[56]

To conclude this section on discrimination and social data, we emphasize the context of these findings. We have shown that computers are susceptible to producing biased decisions, in a broader variety of contexts than most would imagine. Putting signaling theory together with researchers' findings on how algorithms detect patterns around social categories, we have shown that computers can discriminate whether or not they are directed to and whether or not they are given social category information. Algorithmic systems could even introduce arbitrary, self-perpetuating bias, though of course bias is more likely when the data reflect existing prejudice. In focusing on these edge cases, we are not dismissing or denying other processes behind discrimination. In fact, other economists' empirical results remind us that racial discrimination is part of the measured reality of the U.S. labor market and will be reflected in big data.[57]

In this next section, however, we aim to provide hope and work to address these dilemmas. We argue that detecting and alleviating patterns of discrimination is possible when

[56] Varshney and Varshney. While they apply their model to human decision-makers, such as police officers on patrol, some algorithms might implement similar processes, especially in situations where there is limited data available about some subgroups. Indeed, Yee and Ho apply a similar principle to identify and fix problems with comparing discretized test score distributions.

[57] See, for instance, recent economic evidence from Charles and Guryan and Fryer, Pager, and Spenkuch. Both find some support for Becker's model of *taste based discrimination*, in which local employers' racist attitudes affect wages. Goldberg presents an alternate economic model where arbitrary bias can persist if sustained by nepotism.

social category data are available. Moreover, in order to have a deep understanding of social problems, it is imperative that sensitive social category information be available.

## Using Social Category Data: Scenarios, Proposed Remedies, and Risks

We have reviewed theory and evidence showing that removing sensitive categories does not prevent algorithmic discrimination. In the following section, we examine situations and strategies where outcomes improve by considering sensitive categories. Each scenario begins with a case in which using data labeled with social categories allows researchers to detect, understand, or remediate patterns of discrimination. It then describes additional strategies in the same spirit, many of which are more specifically targeted to algorithmic decision-making. In these scenarios, goals would likely not be reached without sensitive social category information.

### *Case 1: Using Data for an External Audit*

In multi-year processes like primary and secondary schooling, knowing when differences arise is an important step to overcoming those differences and preventing them from becoming more extreme. Education researchers often act as watchdogs, comparing how students' school experiences compare to past benchmarks and comparing these student experiences across different social categories; often their findings shine a light on inequities and contribute to innovations in theory, policy, and practice. Here we show how analyses in this spirit have documented skill gaps upon entering school and systematic inaccuracies in teachers' assessments of student ability. We then introduce the literature about how to audit algorithmic systems.

Reardon and Portilla analyze school readiness upon kindergarten entry, using longitudinal studies of U.S. students. These data include the sensitive data of family income and race/ethnicity, as well as professionally-administered cognitive assessments and reports of the

children's behavior and experiences by parents and teachers. ~~Thus~~Thus, the researchers can track skill gaps by income and by race/ethnicity in student measures over time. They find that the gap between white and Hispanic children in average school readiness has narrowed from 1998 through 2010. Relatedly, they also find that the school readiness gap across children from different income brackets decreased over the same time period.[58]

Ready and Wright use one of the same datasets to compare how kindergarten teachers and external assessors rate each student's cognitive ability. They find that many teachers tend to rate girls, white children, and children of higher social class as having higher literacy skills than their classmates. "[A]pproximately half of the sociodemographic disparities in teacher perceptions … are rooted in reality,"[59] as confirmed by those students' independent literacy test scores, while the rest appears to be systematic error by group. Further, teachers in "higher-achieving and higher-SES classrooms" tend to overestimate all children's abilities, while those in "lower-achieving and socioeconomically disadvantaged classrooms" systematically underestimate their students.[60] Together, these studies remind us that differences by social categories are pervasive, and that social category data are necessary for uncovering some of these hidden forces.

***Remedy 1: Auditing Algorithms and Data Systems***

This watchdog model, where researchers look at official data to measure bias in schools, is parallel to the external auditing paradigm. Sandvig and colleagues suggest how systematic

[58] Reardon and Portilla.

[59] Ready and Wright, 348.

[60] Ibid., 351.

"algorithmic auditing" of online services can be accomplished. They review the kinds of audits that might be used, from code reviews (which may not be valuable without also seeing the data), surveying consumers about their background information and experiences with the services, a data scraping audit (which might violate the terms of service and the law), a "sock puppet" audit (using a computer to make false profiles to test the system, though this could violate the law and terms of service), and a "crowdsourced" or "collaborative" audit, where many human testers are recruited to do a systematic audit.[61] They point out key research questions that could support our ability to fairly audit algorithms: "How difficult is it to audit a platform by injecting data without perturbing the platform? What is the minimum amount of data that would be required to detect a significant bias in an important algorithm? What proofs or certifications of algorithmic behavior could be brought to bear on public interest problems of discrimination?"[62]

After-the-fact business audits may also be possible. Citron advocates that a government agency audit private companies' algorithms,[63] and Ramirez, as former head of the Federal Trade Commission (FTC), argues that "[at] the very least, companies must ensure that by using big data algorithms they are not accidentally classifying people based on categories that society has decided—by law or ethics—not to use, such as race, ethnic background, gender, and sexual orientation."[64] For instance, hotel-alternative Airbnb recently set up a program of racial discrimination audits as part of an investigation by the California Department of Fair

[61] Sandvig et al., 12-15.

[62] Ibid., 18.

[63] Citron.

[64] Ramirez, 8.

Employment and Housing.[65] With governmental or corporate support and sufficient resources, institutionalized audits might be used to detect bias in many contexts. If not given such access, researchers may need to consider the methods of external auditing that are still under development by Sandvig and colleagues.

***Case 2: Detecting and Removing Biased Questions on Standardized Tests***

Another approach recognizes that creators of measures and algorithms often want to be fair. This section begins with the concrete example of detecting and removing biased questions on standardized tests. From there, we survey different approaches to tools, standards, and self-governance for people creating and deploying algorithms.

People who develop standardized tests today—usually psychometricians—recognize that their work, and the scores it produces, will shape other people's lives. Early standardized tests often asked questions imbued with class, race, and gender biases. Some early test-designers tried to address this problem, while others believed in eugenics and embraced these biases.[66] Today, the questions included in high-stakes tests are evaluated for *differential item functioning*: typically, a new test item is added, amidst existing test items, and experts reject it if test takers with the same scores everywhere else perform differently on it by demographic subgroup.[67] This process can catch items where wording has additional connotations to some groups of test-takers or where other small differences flag social identity rather than the skills the test is intended to measure.

[65] Guynn.

[66] See Camilli, 235, 249.

[67] Ibid., 226.

In fact, psychometricians think carefully about the uses for which their assessments would be valid, and the circumstances in which they would not be valid. Part of confirming this validity is rigorously comparing the measures they have to the outcomes their tests claim to predict and measures of related traits. Establishing validity also requires checking whether the inferred correspondences hold across different subgroups of people.[68] In the same way, when algorithmic judgment is used in high stakes situations, the creators and users of those algorithms should aim to validate scores there too.

### *Remedy 2: Internal Testing to Build Better Algorithms*

Romei and Ruggieri provide an excellent multidisciplinary and multinational review of data-based discrimination. They first define different discrimination phenomena (from favoritism to tokenism) and summarize how discrimination is treated in common law countries and the European Union. They discuss not only social scientific models of detecting and stopping discrimination, but also models that are oriented towards scholars of data mining and knowledge discovery in databases (KDD). In particular, creators of algorithmic systems have three general classes of approaches to prevent discrimination: they can make the data less biased beforehand, build fairness criteria into the algorithm, or alter the application of the rules after the algorithm runs.

One such strategy, exploratory discrimination-aware data mining (DADM), identifies potential discrimination that should undergo further analysis.[69] Such strategies are being

[68] Ibid., 227, 230.

[69] Berendt and Preibusch.

developed and advanced by the Fairness, Accuracy, and Transparency in Machine Learning (FAT/ML) research community. Similarly, the Association for Computing Machinery this year released a list of seven standards for making algorithms fairer.[70] When the creators and users of data-driven judgments are committed to avoiding unjust bias, more options are available for preventing algorithmic discrimination. Unfortunately, best practices and methodology are still developing, and they will continue to do so as new computational methods develop.

***Case 3: Recognizing How Legal, Ethical, and Social Concerns Shape Policies and Results***

Even when all entities involved would like to be fair, algorithms—especially if they are part of large sociotechnical systems—are often shaped by competing values. Algorithms usually seek the optimal value for a central outcome of interest; for instance, they might aim to predict how productive each job candidate would be if employed and to select the person with the top score. However, they may have additional goals to satisfy, shortcuts to reduce processing time, or other constraints written in. For example, Pope and Sydnor examined a state unemployment program that mandated workshop attendance for people predicted to exhaust their unemployment benefits without finding a new job; although the program had data about workers' race, age, sex, and citizenship, it was prohibited from using them in the assignment process.[71] The authors find that different people would receive this workshop if the model considered social category data.

Simulating who would have been required to attend the workshop under different decision rules, the authors find significant omitted variables bias from excluding race, sex, age,

[70] Association for Computing Machinery US Public Policy Office.

[71] Pope and Sydnor, 219.

and citizenship from the model. Different sets of predictors of employment status would have been selected if these variables had been included. For example, they found that a construction industry indicator was a decent proxy for race and that including age in the model changes how much tenure (i.e. duration of present employment) matters.[72]

Two factors relevant to policymakers are in tension in this case. On the one hand, having "accurate predictions of the outcome of interest"[73] is beneficial, which happens when included variables can freely act as proxies for excluded variables. On the other hand, it is also good is to "giv[e] appropriate relative weight to the different predictors in the model."[74] When omitted variable bias distorts the relationship between predictors and the desired outcome, it subtly bakes forbidden variables into the decision anyway, and it also encourages individuals to treat the distorted predictors as signals that they can manipulate. In the end, Pope and Sydnor advocate for keeping social category information in prediction models to prevent omitted variables bias, but to drop social category information from the subsequent step of scoring individuals.

***Remedy 3: Recognizing Trade-offs Between Desired Outcomes Identifies the Solution Space***

There are often trade-offs in trying to create accurate, discrimination-free models, especially if there are constraints on which factors can be used. In many circumstances, U.S. courts have found it illegal to make an employment decision based on characteristics of a protected group through statistical discrimination, regardless of whether the employer's beliefs

[72] Ibid., 222-223.

[73] Ibid., 217.

[74] Ibid., 217.

about that group are accurate.[75] However, a "business necessity" loophole may permit some data mining algorithms to statistically discriminate anyway.[76] Put another way, even when explicit discrimination is illegal, some data-driven biases may be legally justifiable. When the algorithms update in response to new observations, the situation may be even harder to regulate: European Union safeguards against discrimination during border crossings, designed to address static decision rules, may be unable to address constantly self-updating profiling algorithms.[77]

Beyond the explicitly illegal, we also wish to consider what is ethical. Different definitions of fairness have been proposed for this problem, and in almost all situations, they require tradeoffs between different principles. Three key frameworks are provided by Dwork and colleagues; Friedler, Scheidegger, and Venkatasubramanian; and Kleinberg, Mullainathan, and Raghavan. None of the frameworks excuses us from further responsibility for what our algorithms do: compromising one value to address another is almost always necessary, since solutions matching all of the desired criteria for fairness are unlikely to exist in most cases, and finding the answer, even if it existed, would be extremely hard.[78]

### *Case 4: Inclusion in STEM Fields Requires Focus on Underrepresented People*

[75] Akerlof and Kranton, 91.

[76] Barocas and Selbst, "Big Data's Disparate Impact."

[77] Leese.

[78] Kleinberg, Mullainathan, and Raghavan ("Inherent Trade-Offs in the Fair Determination of Risk Scores") prove that adding a common-sense constraint to their fairness definitions would make the problem *NP-complete*. That computer science classification means that solutions must be found case by case, without any currently known efficient solution.

The previous approaches implicitly discount the distinctiveness of particular social categories. In essence, the above models are built around the experience of the statistically "typical" person in the data, measuring everyone else against those terms. However, the model may not fit correctly, and it may fail systematically for certain groups. Dwork provides an example relevant to bias in hiring in science, technology, engineering, and math (STEM) jobs:

> Suppose we have a minority group in which bright students are steered toward studying math, and suppose that in the majority group bright students are steered instead toward finance. An easy way to find good students is to look for students studying finance, and if the minority is small, this simple classification scheme could find most of the bright students. … A true understanding of who should be considered similar for a particular classification task requires knowledge of sensitive attributes, and removing those attributes from consideration can introduce unfairness and harm utility.[79]

Indeed, for both innocuous and suspect reasons, there are systematic differences by social category in higher education for science, technology, engineering, and mathematics (STEM). For instance, Sheppard and colleagues use large, carefully collected datasets to examine the career paths that trained engineers take, finding that decisions are influenced by gender and status as an underrepresented minority, as well as by factors like subfield and working conditions.[80] Lichtenstein and colleagues further address specific findings on race, ethnicity, and gender in

[79] Miller, quoting Dr. Cynthia Dwork.

[80] Sheppard et al.

engineering by reviewing past studies, suggesting policies that schools and employers could implement to mitigate the loss of diverse talent between one career stage and the next.[81]

***Remedy 4: Understand the System as Experienced by a Particular Group***

Important processes work differently across social groups. Hancock dubs many social challenges "causally complex," explaining that "there are multiple causal recipes that sets of individuals can pursue to the same outcome of interest, whether that outcome is dismissal of criminal charges, delay of deportation proceedings, access to proper HIV/AIDS medical treatment, or high school graduation."[82] We might especially expect this causal complexity to develop in a highly socially differentiated system.[83]

Furthermore, where social categories intersect there may be intensified versions or unique forms of discrimination. To introduce the term "intersectionality," Crenshaw argues, "Black women sometimes experience discrimination in ways similar to white women's experiences; sometimes they share very similar experiences with Black men. Yet often they experience double-discrimination… [or] experience discrimination as Black women—not the sum of race

[81] Lichtenstein et al.

[82] Hancock, 277.

[83] Spence's model of multiple signaling equilibria, discussed earlier, is a highly stylized model of one such differentiated system. Many social scientific and historic accounts of segregated or highly stratified societies illustrate how permitted actions and ways of life can vary greatly for different groups of people living in essentially the same place and time.

and sex discrimination, but as Black women."[84] Providing several examples of how this latter discrimination arises in ways not usually experienced by white women or black men, asserting that, "These problems of exclusion cannot be solved simply by including Black women within an already established analytical structure."[85]

Thus, foregrounding the experiences of particular groups, rather than forcing them into a Procrustean bed, helps us understand how to overcome bias in big data. Welles advocates strongly for this methodology: instead of trying to evaluate the status of small subgroups by seeing how they show up in an overall analysis, she analyzes their experiences separately. In one case, she zooms in on an "extreme minority [that] would normally get lost in Big Data analytics, wiped away as noise among the statistically average masses."[86] Because the number of observations starts out so large, she can still use statistical techniques, among others, to sift through the data. Overall, her work allows the experience of these minority groups to be better understood and considered in future design.

Leurs and Shepherd point out that other research paradigms could also make big data analyses more responsive to marginalized groups. They advocate for approaches that emphasize the research subjects' or users' perspectives, "dialogically involving informants as knowledge

[84] Crenshaw, 149. She provides examples from labor discrimination cases, in which courts variously deny black female plaintiffs' representativeness of women in sex cases, deny their commonality with male coworkers in race cases, and deny their standing as a class unto themselves, because they experience this nexus of discrimination that the courts could not agree on how to characterize.

[85] Ibid., 140.

[86] Welles, 2.

co-producers or co-researchers who share valuable insights."[87] For example, if a job-finding site worries its algorithms are have discriminatory impact against a particular group, then they might talk with group members about how they interact with the site, learn about their experiences with job-seeking online and offline, and observe how potential employers are interacting with group members' profiles (if at all). To uncover whether there is discrimination occurring and how to address it, research focused on and created in dialogue with varied members of that group can provide a nuanced approach.

Furthermore, critical researchers who consider racism and other power dynamics have sought quantitative methods that do not force an essentialist, static measurement of social categories. Hancock incisively explains how intersectionality can be operationalized in research, and she suggests a different form of a common analytical tool to better capture ambiguous social categories.[88] How we define and treat social categories changes the meaning of analysis and changes what our investigations can detect.

---

[87] Leurs and Shepherd, 225.

[88] Hancock. She illustrates how fuzzy set qualitative comparative analysis gives political scientists a version of one of their typical methods that better accounts for the subtleties of racial experiences and expression. For further critical scholars writing on applying quantitative methods in the social sciences, see Zuberi and Bonilla-Silva, Else-Quest and Hyde on psychology, and Gillborn, Warmington, and Demack on education. For theoretical clarifications that speak to quantitative as well as qualitative research, see McCall within feminism and Choo and Ferree within sociology; both emphasize that analysts should also examine the privileged reference category, rather than taking its processes, norms, and values as a default, and that analysts must consider the complex processes through which social categories interact in a given context, rather than assuming an additive relationship or a particular nesting of one identity within another.

Moreover, an individual's (self-proclaimed or perceived) membership in a particular social category and the societal position attached to that category vary over time and by context.[89] For instance, Doleac and Hansen see regional variation in ensuing racial bias among young, low-skilled men. They surmise that "employers are less likely to use race as a proxy for criminality in areas where the minority population of interest is larger—perhaps because discriminating against that entire set of job applicants is simply infeasible."[90] Just because different parts of the country may differ in exactly how they discriminate does not mean that the discrimination is not real and important, even if it is harder to measure.

Even more difficult, the identifiers involved might be extremely complex. In her research on students in for-profit colleges, Cottom finds that people with various intersecting marginalized social categories are overrepresented. Much of the marketing targeted people who were vulnerable—for instance, those searching the Internet for "unemployment insurance"—and people in marginalized social categories were more likely to be in that position.[91] She thus urges researchers to recognize that looking for power inequalities may be a better strategy than looking for social markers. The particular social markers linked to people who lack power in a situation

[89] Hulko.

[90] Doleac and Hansen, 5. Their study, like Agan and Starr's, considers the impact of "Ban the Box"—laws preventing job applications from asking about criminal convictions—on racial discrimination in hiring. In particular, they see that the policy raises discrimination against young black men everywhere except in the U.S. South—the region with the highest black population. Similar results appear to hold for young Hispanic men in the U.S. West—again, the region with the highest Hispanic population—although the relevant results are not all statistically significant.

[91] Cottom.

might change over time, but the power inequalities that are part of certain practices and institutions endure.

Within Internet studies, informatics, and related fields, consideration of social categories and discrimination is unfortunately rare.[92] In a review on racism specifically for the field of Internet studies, Daniels reminds readers that "the preponderance of research about the Internet done by white people … rarely acknowledges the salience of race but instead clings to the fantasy of a color-blind web."[93] Robinson and colleagues provide an excellent analysis of digital inequality, including documenting the processes by which digital inequality interacts with gender, race, and class.[94] In response to evidence that ICT use and various outcomes may differ by social category, it is vital for researchers to seek out the mechanisms and processes which produce these different patterns of ICT use and digital traces.

In short, the education and employment examples here provide substantial evidence that sensitive data can be crucial for understanding biases; the strategies they are paired with provide promising inroads toward countering algorithmic discrimination. Sensitive data unearth both statistical patterns and causal relationships. Decisions based on these data can generate unfair outcomes, but if done thoughtfully they can prevent discrimination. Though we argue for the need to collect social identifiers and have provided a set of cases to make that assertion, in the next section we elaborate on some of the risks inherent in personal data collection.

---

[92] Gandy and Nakamura each provided early discussions of race online, and they have continued research in various facets.

[93] Daniels, 720.

[94] Robinson ~~et al.~~et al. Also, see Gilbert for a critical review of the more typical "digital divide" literature.

## Risks of Data Collection

Unfortunately, while uncovering bias may require collecting sensitive identity data, doing so can also entail potential harms and risks. General concerns about data protection and privacy are heightened when respondents have marginalized identities or identities that will be relatively rare in that context. Their social categories increase the likelihood that their data can be de-anonymized, the risk that they will be targeted specifically, or the harm that could occur if their records are compromised.

Camilli discusses several technical and policy concerns that can arise when collecting and using social category data. Echoing the critical researchers mentioned above, he points out that a social label may group together people who are actually very different on attributes being studied. Furthermore, referring to social categories may lend credence to beliefs in group inferiority or superiority or in ideas of "fixed biological or ethnic classification,"[95] or entrench cumulative disadvantage.[96]

Furthermore, the act of asking about a social category can induce *stereotype threat*, if the respondent has an identity associated with negative stereotypes in that context. Social psychologists have found that asking someone to identify their social category can remind them of negative stereotypes about that part of their identity and of how others may judge them, and this results in worse performance on tests that follow.[97] Luckily, it is possible to reduce the harms of stereotype threat. On a practical level, high-stakes data should not be collected too soon

[95] Camilli, 244.

[96] Gandy.

[97] See Steele and Aronson and Walton et al.

after asking questions meant to elicit a person's social category. More broadly, organizations should consider whether they are perceived as fairly serving people of different social identities. Committing to communications, policies, and actions that respect people across social categories may help avoid stereotype threat by reducing the power of the stereotypes themselves.

## Conclusion

While this paper has provided examples of discrimination in algorithms, determining what constitutes unacceptable judgment is a critical question that should be decided before implementing an algorithm, and revisited in light of the algorithm's evolution and outcomes. Certainly we stand on the frontier of an increasingly large digital landscape, fraught with data-collection dilemmas, privacy concerns, new decision-making standards, and shifting norms in the labor market. This new realm reflects many of the social constructs and problems we have faced before.

Along with our lengthy discussion about algorithmic decision making and discrimination, we have provided a set of remedies tied to four scenarios, and have considered the potential applicability—and drawbacks—of each remedy. While our examples focus on racism in education and employment—all important policy and scholarly domains in themselves—algorithmic discrimination is a threat in many contexts. The remedies we presented for algorithmic bias include external auditing, designing algorithmic judgments to be valid across social categories, considering how legal and other limitations shape algorithmic systems, and focusing research on a particular group in order to gain insight. Each approach is still developing, and the correct strategy to take will depend on the context and researchers' access to the sociotechnical system, its data, and its participants, as well as the analytical tools available to researchers and accepted as evidence in their domain. Our remedies involve the collection of

social category data, and we expect ongoing scholarly dialogue about these proposed solutions. In an effort to begin teasing out some of the dilemmas tied to these scenarios and remedies provided in this work, we add here several deeper questions to spur additional scholarly thinking in this area.

First, we pose a need to analyze feedback effects. Feedback effects can reinforce arbitrary or discriminatory biases, and they are especially pernicious if it is not transparent how "new" inputs for the algorithm stem from prior decisions. Thus we must understand each algorithm's larger context. For example, public school test scores were never intended to be advertised by real estate agents, yet in many parts of the U.S. they strongly influence housing prices.[98] The processes—education measurement and home-buying—are related in a sociotechnical system. A statistical blip in a neighborhood—erroneously low (or high) school test scores, or the departure (or arrival) of several residents who were willing to pay more to live in a place with "good" schools—might become a self-reinforcing trend.

Further, changes in society can ripple through sociotechnical systems, too. Ideas of what is legal, moral, and possible affect how decision-making systems develop, and how people respond to them. For instance, Poon convincingly argues that credit scores were promoted by creditors to forestall the U.S. from creating anti-discrimination regulations for their industry.[99] And, as Stiglitz reminds us, people who can positively differentiate themselves through a market signal tend to do so, *even when creating the signal is not, in itself, useful to them*.[100] Once a

[98] See, for instance, Nguyen-Hoang and Yinger.

[99] Poon.

[100] Stiglitz.

credit score exists, or once ZIP code is a factor in insurance rates, or once a new certification is created, people may change their actions to influence those signals.

Second, scholars and practitioners will no doubt continue to wrangle with epistemological quandaries tied to big data collection. Xie draws our attention to a crucial difference between understanding physical science and "population" science. Researchers analyzing physical phenomena seek universal laws, while those studying living populations—from Darwin to contemporary social scientists—"[recognize] that units of analysis in a population are different from one another, or heterogeneous."[101] To what extent can we assume that similar-seeming people are similar in ways that matter for a given case? He points out that almost any statistical approach requires aggregating information to some extent, and that "we may choose not to analyze (say, by averaging over) within-group, individual-level heterogeneity in a research setting for practical reasons."[102] Understanding this individual variation within a social category is thorny.[103]

Further, we have seen that discrimination plays out differently in different contexts—especially if intersecting social categories are involved. Depending on the scope of data an algorithm learns from, it may or may not replicate this contextual discrimination. But for those checking up on an algorithm, statistical variation may hide relevant issues, unless we comprehend which contexts or which intersectional identities are most sensitive. Complicating this even further, algorithms may synthesize categories that people themselves do not know they

[101] Xie, 6262.

[102] Ibid., 6263.

[103] See Cooper for an extensive discussion of variation within ascribed social categories.

are part of, such as susceptibility to depression;[104] while this may not fall into standard legal frameworks about discrimination, and checking against standard social categories would not detect this, it is an important challenge in algorithmic bias.

Finally, clear analytical results may not lead straightforwardly to a course of action. Social scientists and policymakers often must distinguish between reliable prediction, understanding root causes, and identifying workable policy interventions. Recent work discusses this crucial distinction, in the context of algorithms and automation.[105] Even when an undisputed causal relationship exists, that does not always point to a clear policy lever or action that will work successfully to change course.

Broadly conceived, what we wish to assert with this project is that algorithmic decision-making necessarily relies on correlations. These data relationships may link a person's traits, past actions, social contacts, and social categories to people who were good or bad risks in the past. This process can replicate past discrimination or make assumptions about an individual based on group membership, and it can do so even when an individual keeps data private or when sensitive categories are omitted. This system can create distortions in people's actions as it favors certain signals, and it can potentially magnify chance differences into self-reinforcing discrimination. Algorithms can make predictions from incredibly complex data, but we are ultimately responsible for what they do. Across cultures, contexts, geographic regions, and sociotechnical systems, algorithms must be created with fairness in mind. Then we need to check

[104] See [paper presented at ICA 2017 preconference – probably a submission to this issue].

[105] Kleinberg, Ludwig, Mullainathan, and Obermeyer, "Prediction Policy Problems;" Hoffman, Sharma, and Watts; and Athey.

for unfair outcomes and act to rectify any algorithmic injustices. Ongoing review of how we use algorithmic decision-making and what our algorithms have learned to value over time is critical for a fair society, in any society.

**References**

Agan, Amanda. "Increasing Employment of People with Records." *Criminology & Public Policy* 16, no. 1 (February 2017): 177–85. http://onlinelibrary.wiley.com/doi/10.1111/1745-9133.12266/full.

Agan, Amanda, and Sonja Starr. "Ban the Box, Criminal Records, and Racial Discrimination: A Field Experiment." Working Paper. University of Michigan Law and Economics Research Paper 16-012, 2016. https://28182d77-a-62cb3a1a-s-sites.googlegroups.com/site/amandayagan/Agan%20and%20Starr%2002262017.pdf.

Akerlof, George A., and Rachel E. Kranton. *Identity Economics: How Our Identities Shape Our Work, Wages, and Well-Being*. Princeton, New Jersey: Princeton University Press, 2010.

Anderson, Margo, and William Seltzer. "Federal Statistical Confidentiality and Business Data: Twentieth Century Challenges and Continuing Issues." *Journal of Privacy and Confidentiality* 1, no. 1 (April 14, 2009): 7–52. http://repository.cmu.edu/jpc/vol1/iss1/2.

Angwin, Julia, and Terry Parris, Jr. "Facebook Lets Advertisers Exclude Users by Race." *ProPublica*, October 28, 2016. https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race (accessed May 14, 2017).

Angwin, Julia, Ariana Tobin, and Madeleine Varner. "Facebook (Still) Letting Housing Advertisers Exclude Users by Race." *ProPublica*, November 21, 2017. https://www.propublica.org/article/facebook-advertising-discrimination-housing-race-sex-national-origin (accessed November 22, 2017).

Arrow, Kenneth J. "Some Models of Racial Discrimination in the Labor Market." DTIC Document, 1971. http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=AD0735068 (accessed April 25, 2017).

———. "What Has Economics to Say about Racial Discrimination?" *The Journal of Economic Perspectives* 12, no. 2 (1998): 91–100. http://www.jstor.org/stable/2646963.

Association for Computing Machinery US Public Policy Office. "Statement on Algorithmic Transparency and Accountability," January 12, 2017. https://techpolicy.acm.org/?p=6156 (accessed May 9, 2017).

Athey, Susan. "Beyond Prediction: Using Big Data for Policy Problems." *Science* 355, no. 6324 (February 3, 2017): 483–85. doi:10.1126/science.aal4321.

Barocas, Solon, and Andrew D. Selbst. "Big Data's Disparate Impact." *California Law Review* 104 (2016): 671–732.

———. "Losing Out on Employment Because of Big Data Mining." *The New York Times*, August 6, 2014, sec. Room for Debate: Is Big Data Spreading Inequality? https://www.nytimes.com/roomfordebate/2014/08/06/is-big-data-spreading-inequality/losing-out-on-employment-because-of-big-data-mining (accessed May 18, 2017).

Becker, Gary S. *The Economics of Discrimination*. Second Edition, 2010 Reissue. University of Chicago Press, 1972. http://www.myilibrary.com?ID=273836.

Berendt, Bettina, and Sören Preibusch. "Better Decision Support through Exploratory Discrimination-Aware Data Mining: Foundations and Empirical Evidence." *Artificial Intelligence and Law* 22, no. 2 (June 1, 2014): 175–209. doi:10.1007/s10506-013-9152-0.

Bertrand, Marianne, and Sendhil Mullainathan. "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *The American Economic Review* 94, no. 4 (2004): 991–1013.

Brown, Patricia Leigh. "Creating a Safe Space for California Dreamers." *The New York Times*, February 3, 2017. https://www.nytimes.com/2017/02/03/education/edlife/daca-undocumented-university-of-california-merced-fiat-lux-scholars.html (accessed February 13, 2017).

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan. "Semantics Derived Automatically from Language Corpora Contain Human-like Biases." *Science* 356, no. 6334 (April 14, 2017): 183–86. doi:10.1126/science.aal4230.

Camilli, Gregory. "Test Fairness." In *Educational Measurement*, edited by R. Brennan, 4th ed., 221–256. Westport, CT: American Council on Education and Praeger, 2006. https://www.researchgate.net/profile/Gregory_Camilli/publication/265086461_Test_fairness/links/578e4ae908ae81b4466ec0f8.pdf.

Charles, Kerwin Kofi, and Jonathan Guryan. "Prejudice and Wages: An Empirical Assessment of Becker's The Economics of Discrimination." *Journal of Political Economy* 116, no. 5 (October 1, 2008): 773–809. doi:10.1086/593073.

Choo, Hae Yeon, and Myra Marx Ferree. "Practicing Intersectionality in Sociological Research: A Critical Analysis of Inclusions, Interactions, and Institutions in the Study of Inequalities*." *Sociological Theory* 28, no. 2 (June 1, 2010): 129–49. doi:10.1111/j.1467-9558.2010.01370.x

Citron, Danielle Keats. "Big Data Should Be Regulated by 'Technological Due Process.'" *The New York Times*, August 6, 2014 (updated July 29, 2016), sec. Room for Debate: Is Big Data Spreading Inequality? https://www.nytimes.com/roomfordebate/2014/08/06/is-big-data-spreading-inequality/big-data-should-be-regulated-by-technological-due-process (accessed May 18, 2017).

Cooper, Brittney. "Intersectionality." Edited by Lisa Disch and Mary Hawkesworth. *The Oxford Handbook of Feminist Theory*. Oxford University Press, February 1, 2016. doi: 10.1093/oxfordhb/9780199328581.013.20

Cottom, Tressie McMillan. "Black CyberFeminism: Ways Forward for Intersectionality and Digital Sociology." In *Digital Sociologies*, edited by Karen Gregory, Tressie McMillan Cottom, and Jessie Daniels. Policy Press, 2016.

Coutin, Susan Bibler, Sameer M. Ashar, Jennifer M. Chacón, and Stephen Lee. "Deferred Action and the Discretionary State: Migration, Precarity and Resistance." *Citizenship Studies* 21, no. 8 (2017): 951–68. doi: 10.1080/13621025.2017.1377153.

Crenshaw, Kimberlé. "Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics." *University of Chicago Legal Forum*, 1989, 139–68.

Daniels, Jessie. “Race and Racism in Internet Studies: A Review and Critique.” *New Media & Society* 15, no. 5 (August 1, 2013): 695–719. doi:10.1177/1461444812462849

Doleac, Jennifer L., and Benjamin Hansen. “Does ‘Ban the Box’ Help or Hurt Low-Skilled Workers? Statistical Discrimination and Employment Outcomes When Criminal Histories Are Hidden.” Working Paper. National Bureau of Economic Research, July 2016. http://www.nber.org/papers/w22469.

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. “Fairness Through Awareness.” In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226. ITCS ’12. New York, NY, USA: ACM, 2012. doi:10.1145/2090236.2090255.

Else-Quest, Nicole M., and Janet Shibley Hyde. “Intersectionality in Quantitative Psychological Research: II. Methods and Techniques.” *Psychology of Women Quarterly* 40, no. 3 (September 1, 2016): 319–36. doi:10.1177/0361684316647953

Friedler, Sorelle A., Carlos Scheidegger, and Suresh Venkatasubramanian. “On the (Im)possibility of Fairness.” *arXiv:1609.07236 [Cs, Stat]*, September 23, 2016. http://arxiv.org/abs/1609.07236 (accessed May 9, 2017).

Fryer, Roland G., Devah Pager, and Jörg L. Spenkuch. “Racial Disparities in Job Finding and Offered Wages.” *The Journal of Law and Economics* 56, no. 3 (August 1, 2013): 633–89. doi:10.1086/673323.

Gaddis, S. Michael. "How Black Are Lakisha and Jamal? Racial Perceptions from Names Used in Correspondence Audit Studies." *Sociological Science* 4, no. 19 (2017): 469–489. doi: 10.15195/v4.a19.

Gandy, Oscar H. "Engaging Rational Discrimination: Exploring Reasons for Placing Regulatory Constraints on Decision Support Systems." *Ethics and Information Technology* 12, no. 1 (March 1, 2010): 29–42. doi:10.1007/s10676-009-9198-6

Gilbert, Melissa. "Theorizing Digital and Urban Inequalities." *Information, Communication & Society* 13, no. 7 (October 1, 2010): 1000–1018. doi:10.1080/1369118X.2010.499954

Gillborn, David, Paul Warmington, and Sean Demack. "QuantCrit: Education, Policy, 'Big Data' and Principles for a Critical Race Theory of Statistics." *Race Ethnicity and Education* 0, no. 0 (September 27, 2017): 1–22. doi: 10.1080/13613324.2017.1377417

Goldberg, Matthew S. "Discrimination, Nepotism, and Long-Run Wage Differentials." *The Quarterly Journal of Economics* 97, no. 2 (May 1, 1982): 307–19. doi:10.2307/1880760.

Goldin, Claudia, and Cecilia Rouse. "Orchestrating Impartiality: The Impact of 'Blind' Auditions on Female Musicians." *The American Economic Review* 90, no. 4 (2000): 715–41. http://www.jstor.org/stable/117305.

Griggs v. Duke Power Co. 401 U.S. 424. Supreme Court of the United States. 1971. https://scholar.google.com/scholar_case?case=8655598674229196978.

Guynn, Jessica. "Airbnb to Let California Test for Racist Hosts." *USA Today*, April 28, 2017. https://www.usatoday.com/story/tech/news/2017/04/28/airbnb-let-california-test-racist-

hosts-after-reports-of-bias-against-african-americans/101032640/ (accessed May 1, 2017).

Hancock, Ange-Marie. "Empirical Intersectionality: A Tale of Two Approaches." *UC Irvine Law Review* 3, no. 2 (2013): 259–96.

Heeren, Geoffrey. "The Status of Nonstatus." *American University Law Review* 64, no. 5 (June 2015): 1115–81.

Hofman, Jake M., Amit Sharma, and Duncan J. Watts. "Prediction and Explanation in Social Systems." *Science* 355, no. 6324 (February 3, 2017): 486–88. doi:10.1126/science.aal3856.

Hulko, Wendy. "The Time- and Context-Contingent Nature of Intersectionality and Interlocking Oppressions." *Affilia* 24, no. 1 (February 1, 2009): 44–55. doi:10.1177/0886109908326814

Iversen, Torben, and Frances McCall Rosenbluth. "Explaining Occupational Gender Inequality: Hours Regulation and Statistical Discrimination." Rochester, NY: Social Science Research Network, 2011. https://papers.ssrn.com/abstract=1900012 (accessed April 25, 2017).

Kim, Pauline T. "Data-Driven Discrimination at Work." *William & Mary Law Review* forthcoming (2017). https://papers.ssrn.com/abstract=2801251 (accessed May 1, 2017).

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. "Prediction Policy Problems." *The American Economic Review* 105, no. 5 (May 2015): 491–95. doi:10.1257/aer.p20151023.

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. "Inherent Trade-Offs in the Fair Determination of Risk Scores." In *Proceedings of Innovations in Theoretical Computer Science (ITCS)*, 2017. http://arxiv.org/abs/1609.05807 (accessed May 4, 2017).

Leese, Matthias. "The New Profiling: Algorithms, Black Boxes, and the Failure of Anti-Discriminatory Safeguards in the European Union." *Security Dialogue* 45, no. 5 (October 1, 2014): 494–511. doi: 10.1177/0967010614544204.

Lerman, Jonas. "Big Data and Its Exclusions." *Stanford Law Review Online* 66 (2014 2013): 55–64.

Leurs, Koen, and Tamara Shepherd. "Datafication and Discrimination." In *The Datafied Society: Studying Culture through Data*, edited by Karin van Es and Mirko Tobias Schafer, 211–31. Amsterdam University Press, 2017. http://www.academia.edu/29002676/Datafication_and_discrimination.

Lichtenstein, Gary, Helen L. Chen, Karl A. Smith, and Theresa A. Maldonado. "Retention and Persistence of Women and Minorities along the Engineering Pathway in the United States." In *Cambridge Handbook of Engineering Education Research*, 311–334. Cambridge University Press, 2015.

Lundberg, Shelly J. "The Enforcement of Equal Opportunity Laws Under Imperfect Information: Affirmative Action and Alternatives." *The Quarterly Journal of Economics* 106, no. 1 (February 1991): 309–26. doi:10.2307/2937919.

McCall, Leslie. "The Complexity of Intersectionality." *Signs: Journal of Women in Culture and Society* 30, no. 3 (March 1, 2005): 1771–1800. doi:10.1086/426800

McPherson, Miller, Lynn Smith-Lovin, and James M. Cook. "Birds of a Feather: Homophily in Social Networks." *Annual Review of Sociology* 27 (2001): 415–44. doi:10.1146/annurev.soc.27.1.415.

Miller, Claire Cain. "Algorithms and Bias: Q. and A. With Cynthia Dwork." *The New York Times*, August 10, 2015. https://www.nytimes.com/2015/08/11/upshot/algorithms-and-bias-q-and-a-with-cynthia-dwork.html (accessed May 15, 2017).

Muñoz, Cecilia, Megan Smith, and DJ Patil. "Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights." Washington, DC: Executive Office of the President, The White House, May 2016. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf (accessed May 4, 2017).

Nakamura, Lisa. *Cybertypes: Race, Ethnicity, and Identity on the Internet*. Bristol, PA, USA: Taylor & Francis, Inc., 2002.

"Never Again Tech Pledge," 2016. http://neveragain.tech/ (accessed May 1, 2017).

Nguyen-Hoang, Phuong, and John Yinger. "The Capitalization of School Quality into House Values: A Review." *Journal of Housing Economics* 20, no. 1 (March 1, 2011): 30–48. https://doi.org/10.1016/j.jhe.2011.02.001.

Peppet, Scott R. "Unraveling Privacy: The Personal Prospectus and the Threat of a Full-Disclosure Future." *Northwestern University Law Review* 105 (2011): 1153–1204.

Phelps, Edmund S. "The Statistical Theory of Racism and Sexism." *The American Economic Review* 62, no. 4 (1972): 659–61. http://www.jstor.org/stable/1806107.

Poon, Martha Ann. "What Lenders See—A History of the Fair Isaac Scorecard." Ph.D., University of California, San Diego, 2012. http://search.proquest.com/pqdtglobal/docview/1034339022/abstract/F43B16072B04491DPQ/1.

Pope, Devin G., and Justin R. Sydnor. "Implementing Anti-Discrimination Policies in Statistical Profiling Models." *American Economic Journal: Economic Policy* 3, no. 3 (2011): 206–31. doi:10.1257/pol.3.3.206.

Ramirez, Edith. "The Privacy Challenges Of Big Data: A View From The Lifeguard's Chair." presented at the Technology Policy Institute Aspen Forum, Aspen, Colorado, August 19, 2013. https://www.ftc.gov/public-statements/2013/08/privacy-challenges-big-data-view-lifeguard%E2%80%99s-chair.

Ready, Douglas D., and David L. Wright. "Accuracy and Inaccuracy in Teachers' Perceptions of Young Children's Cognitive Abilities: The Role of Child Background and Classroom

Context." *American Educational Research Journal* 48, no. 2 (April 2011): 335–60. doi:10.3102/0002831210374874.

Reardon, Sean F., and Kendra Bischoff. "Income Inequality and Income Segregation." *American Journal of Sociology* 116, no. 4 (2011): 1092–1153. doi:10.1086/657114.

Reardon, Sean F., and Ximena A. Portilla. "Recent Trends in Income, Racial, and Ethnic School Readiness Gaps at Kindergarten Entry." *AERA Open* 2, no. 3 (July 1, 2016): 2332858416657343. doi:10.1177/2332858416657343.

Robinson, Laura, Shelia R. Cotten, Hiroshi Ono, Anabel Quan-Haase, Gustavo Mesch, Wenhong Chen, Jeremy Schulz, Timothy M. Hale, and Michael J. Stern. "Digital Inequalities and Why They Matter." *Information, Communication & Society* 18, no. 5 (May 4, 2015): 569–82. doi:10.1080/1369118X.2015.1012532

Romei, Andrea, and Salvatore Ruggieri. "A Multidisciplinary Survey on Discrimination Analysis." *The Knowledge Engineering Review* 29, no. 5 (November 2014): 582–638. doi:10.1017/S0269888913000039.

Ross, Stephen L., and John Yinger. *The Color of Credit: Mortgage Discrimination, Research Methodology, and Fair-Lending Enforcement*. The MIT Press, 2002.

Rugh, Jacob S., Len Albright, and Douglas S. Massey. "Race, Space, and Cumulative Disadvantage: A Case Study of the Subprime Lending Collapse." *Social Problems* 62, no. 2 (May 1, 2015): 186–218. doi:10.1093/socpro/spv002.

Rugh, Jacob S., and Douglas S. Massey. "SEGREGATION IN POST-CIVIL RIGHTS AMERICA: Stalled Integration or End of the Segregated Century?" *Du Bois Review: Social Science Research on Race* 11, no. 2 (October 2014): 205–32. doi:10.1017/S1742058X13000180.

Sandvig, Christian, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. "Auditing Algorithms." Seattle, WA, USA, 2014. http://www-personal.umich.edu/~csandvig/research/Auditing%20Algorithms%20--%20Sandvig%20--%20ICA%202014%20Data%20and%20Discrimination%20Preconference.pdf (accessed March 24, 2017).

Sheppard, Sheri D., Anthony Lising Antonio, Samantha R. Brunhaver, and Shannon K. Gilmartin. "Studying the Career Pathways of Engineers: An Illustration with Two Data Sets." In *Cambridge Handbook of Engineering Education Research*, 283–310. Cambridge University Press, 2015.

Smith, Edward J., and Shaun R. Harper. "Disproportionate Impact of K-12 School Suspension and Expulsion on Black Students in Southern States." Philadelphia: University of Pennsylvania, Center for the Study of Race and Equity in Education, 2015. http://www.gse.upenn.edu/equity/SouthernStates (accessed May 16, 2017).

Spence, Michael. "Job Market Signaling." *The Quarterly Journal of Economics* 87, no. 3 (August 1, 1973): 355–74. doi:10.2307/1882010.

———. "Signaling in Retrospect and the Informational Structure of Markets." *The American Economic Review* 92, no. 3 (June 2002): 434–59. http://www.jstor.org/stable/3083350.

Steele, Claude M., and Joshua Aronson. "Stereotype Threat and the Intellectual Test Performance of African Americans." *Journal of Personality and Social Psychology* 69, no. 5 (1995): 797–811. doi:10.1037/0022-3514.69.5.797.

Stiglitz, Joseph E. "Information and the Change in the Paradigm in Economics." *The American Economic Review* 92, no. 3 (2002): 460–501. http://www.jstor.org/stable/3083351.

Strahilevitz, Lior Jacob. "Privacy versus Antidiscrimination." *The University of Chicago Law Review* 75, no. 1 (Winter 2008): 363–81. http://www.jstor.org/stable/20141912.

Sweeney, Latanya. "Discrimination in Online Ad Delivery." *Communications of the ACM* 56, no. 5 (2013): 44–54. doi:10.1145/2447976.2447990.

Varshney, L. R., and K. R. Varshney. "Decision Making With Quantized Priors Leads to Discrimination." *Proceedings of the IEEE* 105, no. 2 (February 2017): 241–55. doi:10.1109/JPROC.2016.2608741.

Walton, Gregory M., D. Paunesku, and C. S. Dweck. "Expandable Selves." In *The Handbook of Self and Identity*, edited by M.R. Leary and J.P. Tangney, Second Edition, 141–154. New York: Taylor and Francis, 2012.

Welles, Brooke Foucault. "On Minorities and Outliers: The Case for Making Big Data Small." *Big Data & Society* 1, no. 1 (July 10, 2014): 2053951714540613. doi:10.1177/2053951714540613.

Xie, Yu. "Population Heterogeneity and Causal Inference." *Proceedings of the National Academy of Sciences* 110, no. 16 (2013): 6262–68. doi:10.1073/pnas.1303102110.

Yee, Darrick, and Andrew Ho. “Discreteness Causes Bias in Percentage-Based Comparisons: A Case Study From Educational Testing.” *The American Statistician* 69, no. 3 (July 3, 2015): 174–81. doi: 10.1080/00031305.2015.1031828.

Žliobaitė, Indrė, and Bart Custers. “Using Sensitive Personal Data May Be Necessary for Avoiding Discrimination in Data-Driven Decision Models.” *Artificial Intelligence and Law* 24, no. 2 (June 1, 2016): 183–201. doi:10.1007/s10506-016-9182-5.

Zuberi, Tufuku, and Eduardo Bonilla-Silva. *White Logic, White Methods: Racism and Methodology*. Lanham, MD, USA: Roman & Littlefield, 2008.