How Algorithms Discriminate Based on Data They Lack:

Challenges, Solutions, and Policy Implications

Betsy Williams	Catherine Brooks	Yotam Shmargad
School of Information	School of Information	School of Information
University of Arizona	University of Arizona	University of Arizona

Abstract

In order to protect consumer privacy and prevent discrimination, companies sometimes choose to delete or resist collecting data about people's social categories (e.g. gender, race). We argue that such censoring can lead to more discrimination by making biases more difficult to detect. We begin with a detailed explanation of how computerized decisions can be biased, even in the absence of social category data. We then show how proactively using social category data can illuminate and combat discriminatory practices, relying on cases from the field of education. We conclude with strategies for detecting and preventing discrimination, and implications for researchers and policymakers.

The Never Again Tech Pledge, drafted and signed by employees of United States tech companies, encourages people to "refuse to participate in the creation of [government] databases … to target individuals based on race, religion, or national origin."[1] Signatories of the pledge further vow to minimize sensitive data collection and "to scale back existing datasets with unnecessary racial, ethnic, and national origin data." These people rightly worry about the government using such data to aid in mass deportations and the internment of immigrants or Muslims.

Data about social categories are powerful, and can certainly be wrongly used to harm. While the many risks of data collection and storage are well-known, there are also issues that 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 marginalized groups.

This paper addresses questions that information professionals, policy-makers, scholars, and Never Again Tech Pledge supporters wrestle with. When is it appropriate to collect and use sensitive information? When does it cause harm, and when does it prevent harm? In this paper, we provide a detailed explanation of how computerized decisions can be biased, including through processes known as *statistical discrimination* and *signaling.* We explain how this can happen even in the absence of social category data. We then show how social category data can

[1] For the full pledge and list of signatories, see http://neveragain.tech/, accessed May 1, 2017.

be used proactively to both illuminate and combat discriminatory practices, taking examples from the education literature. We conclude with strategies for detecting and preventing statistical discrimination, and implications for researchers and policymakers.

Social Identifiers in Big Data

The former Federal Trade Commission Chair, Edith Ramirez, discusses how "big data" are "assembled, bit-by-bit, from little data,"[2] 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."[3]

There is not always agreement about which data might be deemed sensitive, but 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, gender, sexual orientation, 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).

In the United States, many people and institutions still discriminate due to prejudice, and still more make decisions shaped by past prejudices. For example, one of the thorniest social

[2] Ramirez, 4.

[3] Ramirez, 7.

problems that America faces is why important life outcomes—from educational attainment and income to incarceration and life expectancy—systematically vary by race. Many explanations have been proposed, often implicating historical laws and ongoing discriminatory practices. Rugh, Albright, and Massey detail how black Baltimore homeowners were differentially targeted with predatory mortgage loans and faced greater losses in the Great Recession than their white counterparts in similar financial situations.[4] In the labor market, audits continue to show that equivalent candidates are less likely to be interviewed for a job if their names suggest that they are black rather than white.[5]

Because of America's past and present discrimination, data about race, ethnicity, immigration status, and gender are related to many other aspects of a person's life. Where one lives in America remains strongly connected to race[6] and socioeconomic status.[7] Interaction partners also tend to reflect one's social categories as well as other traits.[8] Increasingly, these bits of information about our lives, which are indicative of the many social categories to which we belong, 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

[4] Rugh, Albright, and Massey.

[5] Bertrand and Mullainathan; Agan and Starr.

[6] Rugh and Massey.

[7] Reardon and Bischoff.

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

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 spending helps predict this year's spending behavior). Patterns can 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 are meaningful, and even fewer of which reflect relationships that are causal in nature.

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 a group of 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 that their car was made, their annual income, a sport 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) being 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 have a similar age and income, their cars are more likely to be of the same age. By looking through these sheets we might recognize a few different patterns 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 buy a new car, we could look at the pages for people who recently purchased a car. We could then ask whether there are patterns that appear more often among new car owners. We might then predict that people with patterns similar to new car owners, but who have not recently purchased a new car, are likely to buy a car soon.[9]

[9] With intuition or data over time, we might discern other car-buying patterns: does leasing a car ever lead to buying? Do certain people frequently buy new cars, perhaps trading in their old car every few years? Are there telltale signs that someone will buy a used car, buy their first car, or go carless?

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 for that person by looking at the constellations on other sheets that contain those missing answers and 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 identity or gender. Inevitably, the answer to these questions are going to be associated with lots of other 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 identity, and gender in the data set. We could erase all of the check marks in those columns. However, the constellations that are characteristic to certain answers would still be there. Even if an observer were not intentionally trying to guess, say, gender, they might still see distinctly gendered patterns.[10]

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

[10] Even this example of gender suggests the complexity of why attributes and behaviors may be correlated. Some things may be directly linked to gender itself (such as buying a dress, which would be far more common among women than men) or to its close correlate, sex (such as buying tampons or visiting a gynecologist), while others might have more complex reasons (such as working in a feminized profession or attending a yoga studio).

have later repercussions if laws or contracts change. For instance, if the U.S. changes policies about health insurance or immigration, then sensitive information disclosed under older laws (e.g. pre-existing medical conditions or undocumented immigrant status) could prove detrimental.

For example, through the Deferred Action for Childhood Arrivals (DACA) program, many people registered in order to legally work and receive renewable, two-year protection from deportation. However, their data may still be available even if immigration policies radically shift. Indeed, some young people who put their home addresses in the registry are now experiencing a "'horrific Kafka-like situation' in which they have potentially outed their parents to federal authorities."[11]

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. This is, of course, part of the impulse behind the Never Again Tech Pledge: 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?[12]

When data are "big," unknown data points can be filled in through prediction, imputation, and proxies. Consequently, staying private and holding back personal information cannot always

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

[12] 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.

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,[13] 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 signal.[14] 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?"[15]

There are broader downsides to withholding personal information as well. 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."[16] 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,"[17] 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.

---

[13] Stiglitz.

[14] Peppet.

[15] Ibid., 1159.

[16] Lerman, 56.

[17] Ibid., 57.

Bias and Statistical Discrimination

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. For instance, many major 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. 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.[18]

It may thus feel counterintuitive that leaving out social category data can actually perpetuate discrimination, while collecting and accounting for it can lessen discrimination. Here, we make this very case by discussing four main points. These four points, taken together, suggest that when algorithms use "big data" for important decisions, it is futile—and sometimes harmful—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. 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.*

[18] Goldin and Rouse.

3. 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. Perversely, this means that algorithms can discriminate on the basis of a social category, intentionally and unintentionally, and even when they are not explicitly fed social category data.

4. Detecting and alleviating patterns of discrimination, and ascertaining why they occurred, is only possible when 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.

We elaborate on each of these points in the sections that follow, synthesizing others' work. We distinctly combine these arguments to show how the mechanisms for bias without bigotry fit together with algorithms' prodigious pattern-finding: algorithms using big data, even when social categories are censored, can discriminate based on those categories. After that, we describe several risks that can arise from collecting sensitive data and highlight strategies for exposing and preventing discrimination by algorithms.

Social Identifiers

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., first names of women, men, or both, and occupations often held by women or men). In the same way, the algorithm learned stereotypical biases tied to race and gender, and the authors noted that algorithms that are taught broad associations could be prejudiced in making hiring decisions.[19]

Similarly, Sweeney finds that Google serves different ads depending on the kinds of names that users enter in the search box. For example, ads for background checks were more common for particular races and for males, and 60% of the ads offered for black names mentioned "arrest" or "criminal," versus only 48% for white names.[20] She discusses how companies may have requested these ads and how the differences may have been reinforced.[21]

These studies demonstrate how algorithms can learn negative associations for certain cultural groups, especially when the data reflect a broad array of statistical inputs. This can also arise within an organization's own data.[22] 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."[23] 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

[19] Caliskan, Bryson, and Norayanan.

[20] 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.

[21] Sweeney.

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

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

a feedback loop causes bias in inputs or results of the past to replicate itself in the outputs of an algorithmic system."[24] "Data-driven" decision-making, which these days often relies on big data and sophisticated algorithms, provides plenty of opportunities to associate certain social categories with statistical regularities, stereotypes, and past discrimination.

Statistical Discrimination

Decades ago, labor economists investigated the various mechanisms behind employment discrimination, and one resulting theory is *statistical discrimination*.[25] When employers lack information about an individual job applicant's skills, 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

---

[24] Muñoz, Smith, and Patil, 8.

[25] 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).

extension of this approach, *the signaling model*, shows how such feedback effects can sustain unjustified inferences.[26] Next, we introduce these theories and their evidence, explaining how decades-old models can accurately capture the work of cutting edge algorithms.

Hiring is one of many situations where there is little direct information about how a particular decision will turn out. Bluntly, Arrow posits 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."[27] 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."[28] 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.[29]

---

[26] Spence, "Job Market Signaling."

[27] Arrow, *Some Models*, 21.

[28] 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 Supreme Court case of *Griggs v. Duke Power Co*.

[29] 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.

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 asking up front about criminal convictions.[30] 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.[31] 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."[32]

Statistical discrimination models and studies 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 no

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

[31] The study, still a working paper, uses a careful 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.

[32] Agan and Starr, 28.

different than 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:

> [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.)[33]

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

Arbitrary Discrimination

These papers demonstrate how discrimination can occur without malice. 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*.[34] 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,[35] emphasizing an individual's incentives to develop *signals* of being more productive. We briefly

---

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

[34] 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*).

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

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. Needless to say, if productivity of all workers is the same and is known, wages would be the same for each person. If the workers are indistinguishable, though, then starting wages will tend to be the average of high and low wages, weighted by how many workers of each type are in the market. However, what if there were badges available, which cost low productivity workers twice as much to get as high productivity workers?[36] In pondering the role of such a badge in the marketplace, Spence assumes workers will decide whether to invest in the badge based on its costs and on the wages they would get from this *signal*, and that employers will pay attention to how the badge relates to productivity—the *strength* of the signal.[37]

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.*"[38] 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 multiple different equilibria with different badge prices and hiring outcomes. Indeed, Spence

[36] 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 shows it is likely one was already a high productivity type.

[37] Ibid., 437.

[38] Ibid., 437.

writes that "it is the self-confirming nature of the beliefs that gives rise to the potential presence of *multiple equilibria* in the market."[39]

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 buying a badge, but none of the low productivity workers to do so. If the price of a badge is set somewhere in that range beforehand, then there is no reason for the price of a badge or wages for either group to change, and this is a *separating equilibrium*. However, if enough of the population is high productivity and badges are costly enough, then instead no one will buy a badge and everyone accepts the same wage; this is called a *pooling equilibrium.*[40] 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 and specifies that there are two social category groups, each with the *same mixture of high and low productivity workers.*[41] But even though the situations are symmetric, the situation may evolve differently across the two groups. "One person's signaling

[39] 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: as Kim writes that "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).

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

[41] Spence, "Job Market Signaling," 369.

strategy or decision affects the market data obtained by the employer,"[42] 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, sex—matters. A rational empiricist, the employer now conditions beliefs about productivity on both badges and sex, 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,"[43] and different beliefs—and thus different wages, incentives, and equilibria—can 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 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 them.

[42] Ibid., 370.

[43] Ibid., 370.

Indeed, they demonstrate further conditions under which statistical discrimination can develop without any true between-group differences or malice.[44]

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 we would think. Of course this bias is more likely when data they use reflects people's racism or 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.[45]

Algorithmic Discrimination

Algorithms glean 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 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

[44] Varshney and Varshney. While they apply their model to human decision-makers, some algorithms might implement similar processes, especially in situations where there is limited data available about some subgroups.

[45] 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.

ethnicity. Rather, it guesses a proxy, users' "ethnic 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.[46] If advertisers want to discriminate, this gives them a tool to do so. But even without explicitly developing a proxy variable, omitting social categories can still drive algorithms. We next focus on explaining this problem as it has been discussed in the literature on linear modeling as omitted variable bias. While we focus on the linear case for simplicity, more complex algorithms suffer from the same fundamental problem unless it is explicitly addressed.

Leaving out sensitive variables from an analysis forces correlated variables to take on greater significance. This result is known as *omitted variable bias*. Pope and Sydnor mathematically explore its consequences and 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.[47] When some variables are omitted for being "socially unacceptable for use in predictive models,"[48] their proxies gain heavier use and might themselves become socially unacceptable. For instance, California car insurance rates must not use home

[46] Angwin and Parris, Jr.

[47] 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.

[48] Pope and Sydnor, 210.

locations or credit scores, which too closely reflect the previously-banned factors of race and income.[49]

Pope and Sydnor provide a compelling example by re-analyzing decisions in a U.S. unemployment program. In their setting, unemployed people predicted to exhaust their benefits without finding a new job were required to take a workshop. The authors use three years of data from New Jersey, simulating who would have been required to attend the workshop under different decision rules. They note that while the state had data about workers' race, age, sex, and citizenship, the program was prohibited from using them in the assignment process.[50] They find significant omitted variables bias from excluding race, sex, age, and citizenship from the model: different sets of predictors of employment status would have been selected had they chosen to include these variables in the analysis. For example, they found that a construction industry indicator was a decent proxy for race, and that including age changes the impact of tenure.[51]

Two factors relevant to policymakers are in tension here. On the one hand, having "accurate predictions of the outcome of interest"[52] 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."[53] When omitted variable bias distorts the relationship between predictors and the desired outcome, it subtly bakes

[49] Ibid., 206.

[50] Ibid., 219.

[51] Ibid., 222-223.

[52] Ibid., 217.

[53] Ibid., 217.

forbidden variables into the decision anyway, and it also encourages individuals to treat the distorted predictors as signals that they can manipulate.[54]

This case suggests there are situations where outcomes improve by considering sensitive categories. At the same time, harm is not entirely limited by removing sensitive categories. We have shown that algorithms can discriminate whether or not they are directed to and whether or not they are given social category information. Thus, we next consider the value and risks of collecting social identifiers.

The Benefits and Risks of Collecting Social Category Data

Detecting and thus alleviating patterns of discrimination, and ascertaining why they occurred in the first place, is only possible when data labeled with the social category of interest. Further, deeper understanding of social problems may only be advanced when sensitive social category information is available. To provide evidence of these claims, we draw on cases from education, a field which often collects sensitive identity information and links it to larger datasets. The resulting analyses often have important implications for policy and practice.

Case 1: Detecting and Removing Biased Questions on Standardized Tests

People who develop standardized tests today—usually psychometricians—recognize that their work, and the ratings it produces, shapes 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.[55] Today, the

[54] Ibid., 217-218.

[55] See Camilli, 235, 249.

questions included in such 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.[56] This process may 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.

In fact, psychometricians think carefully about the uses for which their tests 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.[57] When algorithmic judgment is used in high stakes situations, we should aim to validate scores there too. The scoring models should agree with other measures of the same trait and predict relevant outcomes across different segments of the population.

Case 2: Social Categories and Behavior in School

Data about social categories also help practitioners uncover differences that are relevant for policy and practice in ongoing systems like primary and secondary schooling. Our second case documents success gaps upon entering school, systematic inaccuracies in teachers' assessments of student ability, and differing levels of punishment by race.

---

[56] Ibid., 226.

[57] Ibid., 227, 230.

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. Because income and race/ethnicity are included in these data, the researchers can track gaps in student measures over time. They find that the gap between white and Hispanic children in 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 special assessments 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] 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] Smith and Harper similarly document widespread disparities in rates of suspension and expulsion from school by race across the southern U.S. They link it to prior research showing that black students tended to be disciplined for much more subjective

[58] Reardon and Portilla.

[59] Ready and Wright, 348.

[60] Ibid., 351.

behavior—such as "disrespect"—than white students.[61] 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.

Case 3: Social Categories and Developing Talent

In higher education, too, there are systematic differences by social category—for both innocuous and suspect reasons—that should inform the decisions of policymakers. 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.[62] Lichtenstein and colleagues further address specific findings on race, ethnicity, and gender in 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.[63]

Dr. Cynthia Dwork provides a clarifying example about why just such insights need to be accounted for in algorithmically-generated decisions:

> 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

[61] Smith and Harper.

[62] Sheppard et al.

[63] Lichtenstein et al.

classification task requires knowledge of sensitive attributes, and removing those attributes from consideration can introduce unfairness and harm utility.[64]

In short, the education literature provides substantial evidence that sensitive data can be crucial for understanding biases in student evaluation. Sensitive data unearth both statistical patterns and causal relationships. Decisions based on these data can generate unfair outcomes, if done blindly, or it can prevent discrimination, if done thoughtfully. Though we argue for the need to collect social identifiers, we conclude this section my elaborating on some of the risks inherent in personal data collection.

Unfortunately, while uncovering bias may require collecting sensitive identity data, doing so can also entail potential harms and risks. 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, that they would be targeted specifically, and that compromised records will harm them.

Camilli discusses several technical and policy concerns that can arise when collecting and using social category data. He points out that a social label may group together people who are actually very different on attributes being studied, and that referring to social categories may lend credence to beliefs in group inferiority or superiority or in ideas of "fixed biological or ethnic classification."[65]

---

[64] Miller, quoting Dr. Cynthia Dwork.

[65] Camilli, 244.

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.[66]

It is, of course, possible to reduce the harms of stereotype threat. First, high-stakes data should not be collected too soon 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 in each social category may help avoid stereotype threat by reducing the power of the stereotypes themselves.

Strategies for Exposing and Preventing Discrimination by Algorithms

So far, this paper has focused on how data-driven and algorithmic decisions can generate biases, even when sensitive social categorical data are not used. A growing body of literature focuses explicitly on interventions that can be used to expose and prevent such discrimination. We provide a summary of some of these strategies here.

Building 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

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

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, they point to work on preventing discrimination by making the data less biased beforehand, by building fairness criteria into the algorithm, and by altering the rules after the algorithm runs.

One such strategy is called exploratory discrimination-aware data mining (DADM), and identifies potential discrimination that should undergo further analysis.[67] Such strategies are being advanced by the Fairness, Accuracy, and Transparency in Machine Learning (FAT/ML) research community, which is holding its fourth annual conference this year. From another part of the research community, the Association for Computing Machinery this year released a list of seven standards for making algorithms fairer.[68]

External Auditing

While the previous approaches have looked at how the algorithm creators can address discrimination, another important approach is auditing. Sandvig and colleagues suggest how systematic "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

[67] Berendt and Preibusch.

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

many human testers are recruited to do a systematic audit.[69] 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?"[70]

After-the fact business audits may also be possible, with government backing. 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 Employment and Housing.[71] With legal or corporate support and sufficient resources, institutionalized audits can be used to detect bias in online services.

Similarly, Edith 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."[72] Citron agrees, advocating that a government agency audit private companies' algorithms. They would feed data in and see if "algorithmic predictions are statistical proxies for race, gender, religion and disability."[73] She provides as an example that the FTC penalized a subprime credit card

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

[70] Ibid., 18.

[71] Guynn.

[72] Ramirez, 8.

[73] Citron.

company that set consumers' credit limits based on how they were spending their money and the inferences they drew from that about personality.

Research Centering on Specific Groups

Another way of overcoming bias in big data is to conduct studies specifically targeting a subpopulation of interest, as Welles demonstrates. Instead of trying to evaluate the status of small subgroups by seeing how they show up in an overall analysis, Welles analyzes their experience 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."[74] This allows the experience of these minority groups to be fully 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 co-producers or co-researchers who share valuable insights."[75] For example, if a job-finding site might be discriminating against a particular group, then understanding how group members interact with the site and with alternative systems and processes for finding jobs could help the site document the problem while looking for potential solutions.

Practical Implications for Researchers and Policymakers

[74] Welles, 2.

[75] Leurs and Shepherd, 225.

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, and we bring with us many of the social constructs and problems we have faced before. In this final section, we offer a set of practical considerations applicable for researchers, policymakers, and other interested readers.

Legal and Ethical Concerns

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 about that group are accurate.[76] However, there is a business necessity loophole, which some data mining algorithms might use.[77] Put another way, though explicit discrimination is illegal, the data-driven and algorithmic biases we discuss may be legally justifiable, in some cases, and are likely prevalent across industries like advertising, tourism, and banking.

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 schema are provided by Dwork and colleagues; Friedler, Scheidegger, and Venkatasubramanian; and Kleinberg, Mullainathan, and

[76] Akerlof and Kranton, 91.

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

Raghavan. Each of these schemas involves tradeoffs, 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]

Feedback in Sociotechnical Systems

Determinations 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.[79] 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 itself useful to them*.[80] Once a credit score exists, or once ZIP code is a factor in insurance rates, people may change their actions to influence those signals.

There is also a deeper question about what we can learn from data. Xie draws our attention to an epistemological difference between 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."[81] To what extent can we

---

[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.

[79] Poon.

[80] Stiglitz.

[81] Xie, 6262.

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."[82]

Furthermore, 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.[83] However, even when an undisputed causal relationship exists, that does not always point to a clear policy lever to successfully change the situation.

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. If we use an algorithm to help make decisions, we must decide what is fair in that context, create the algorithm with fairness concerns in mind, check for

[82] Ibid., 6263.

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

unfair outcomes, and act to rectify any injustices. Ongoing review of what our algorithms have learned to value is critical for a fair 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*, 2017. https://28182d77-a-62cb3a1a-s-sites.googlegroups.com/site/amandayagan/Agan%20and%20Starr%2002262017.pdf (accessed April 12, 2017).

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).

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.

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).

Dutton, William H., and Kenneth L. Kraemer. “Automating Bias.” *Society* 17, no. 2 (January 1, 1980): 36–41. doi:10.1007/BF02700058.

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.

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.

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).

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.

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).

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.

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).

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

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.

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.

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.