Awareness in Practice: Tensions in Access to Sensitive Attribute Data for Antidiscrimination

Miranda Bogen[*]
Upturn
mirandabogen@gmail.com

Aaron Rieke
Upturn
aaron@upturn.org

Shazeda Ahmed[†]
University of California, Berkeley
shazeda@ischool.berkeley.edu

ABSTRACT

Organizations cannot address demographic disparities that they cannot see. Recent research on machine learning and fairness has emphasized that awareness of sensitive attributes, such as race and sex, is critical to the development of interventions. However, on the ground, the existence of these data cannot be taken for granted.

This paper uses the domains of employment, credit, and healthcare in the United States to surface conditions that have shaped the availability of sensitive attribute data. For each domain, we describe how and when private companies collect or infer sensitive attribute data for antidiscrimination purposes. An inconsistent story emerges: Some companies are required by law to collect sensitive attribute data, while others are prohibited from doing so. Still others, in the absence of legal mandates, have determined that collection and imputation of these data are appropriate to address disparities.

This story has important implications for fairness research and its future applications. If companies that mediate access to life opportunities are unable or hesitant to collect or infer sensitive attribute data, then proposed techniques to detect and mitigate bias in machine learning models might never be implemented outside the lab. We conclude that today's legal requirements and corporate practices, while highly inconsistent across domains, offer lessons for how to approach the collection and inference of sensitive data in appropriate circumstances. We urge stakeholders, including machine learning practitioners, to actively help chart a path forward that takes both policy goals and technical needs into account.

[*]Author was affiliated with Upturn at time of writing.
[†]Author was a Fellow at Upturn at time of writing.

1 INTRODUCTION

Statistical models, including those created with machine learning, can reproduce biases in the historical data used to train them. As powerful institutions increase their reliance upon these models to automate decisions that affect people's rights and life opportunities, researchers have begun developing new techniques to help detect and address these biases. The real-world implementation of these techniques could be an essential part of ensuring the continued viability of civil and human rights protections.

Many machine learning fairness practitioners rely on awareness of sensitive attributes—that is, access to labeled data about people's race, ethnicity, sex, or similar demographic characteristics—to test the efficacy of debiasing techniques or directly implement fairness interventions. A significant body of research presumes the modeler has ready access to data on these characteristics as they build and test their models [13, 19, 21]. The need for this data is plain to see. As a 2003 analysis of racial disparities in healthcare powerfully concluded: "The presence of data on race and ethnicity does not, in and of itself, guarantee any subsequent actions ... to identify disparities or any actions to reduce or eliminate disparities that are found. The absence of data, however, essentially guarantees that none of those actions will occur." [23]

Increasingly, companies that utilize machine learning are being asked to detect and address bias in their products. But they are not the first to grapple with these issues. This paper explores the legal and institutional norms surrounding the collection, inference, and use of sensitive attribute data in three key corporate domains. This analysis has significant implications for machine learning fairness research: If private institutions that mediate access to life opportunities are unable or hesitant to collect or infer sensitive attribute data, then emerging awareness-based techniques to detect and mitigate bias in machine learning models might never be implementable in real-world settings.

Notably, this paper does not discuss complex and important questions about *how* "fairness" should be measured or addressed, recognizing that definitions are manifold [53]. Rather, we make a simpler point: If sensitive attribute data are not available, interventions that rely on them will be severely impaired.

We conduct this exploration through the lens of U.S. civil rights law in the domains of credit, employment, and healthcare. For each domain, we describe when and how private companies collect or infer sensitive attribute data to pursue antidiscrimination goals. These are not the only contexts where collection of sensitive attributes is likely to be justified or important, but they are quintessential areas where such data are already being used to measure and mitigate discrimination. They also highlight major divergences in policy, motivation, and practice.

Comparing these sectors, a complex and inconsistent story emerges. In credit, the law requires some lenders to collect sensitive attribute data, while largely prohibiting others from doing so. In employment, the collection of sensitive attribute data is a familiar part of large employers' day-to-day practice. And in health care, companies' motivation for collecting sensitive attribute data is not just basic antidiscrimination compliance, but rather a moral imperative to address staggering disparities in health outcomes.

We observe that these norms and practices, divergent as they are, typically extend only to traditionally regulated actors. Technology companies that mediate access to opportunities as platforms (e.g., social networks, job boards, and rental sites) or act as vendors to other companies rarely receive clear guidance about when to collect or infer sensitive attribute data. As a result, today, many major technology companies do not collect or infer certain kinds of sensitive attribute data and may therefore struggle to define, detect, and address harms to those protected groups.

We conclude that there are few clear, generally accepted principles about when and why companies should collect sensitive attribute data for antidiscrimination purposes. We emphasize the importance of the machine learning research community engaging on the future development of policy in this area, and urge conversations among stakeholders about whether and how to adapt existing practices or establish new ones.

1.1 Defining "sensitive attribute data"

Throughout this paper, we use the term "sensitive attribute data" to refer to details about people's membership in "protected classes" as defined throughout U.S. civil rights laws. This approach to classification is not without its problems: Rigid categories such as these do not currently accommodate nonbinary identities or membership across multiple groups [3, 34, 37, 51]. We acknowledge the reductive and potentially harmful nature of these classification regimes, while simultaneously emphasizing the importance of understanding how they have motivated data collection practices for bias mitigation, and how the history of these practices can inform contemporary contexts.

1.2 Related work

The fair ML research community has long reflected on the social and policy contexts of its work, recognizing legal tensions [8], historical parallels in prior debates over definitions of fairness [37], and the limitations of data-dependent problem formulation [64, 70]. However, when emphasizing the importance of awareness of sensitive attributes in developing and implementing fairness-enhancing interventions, fair ML research and toolkits [9, 38, 67] often take for granted when framing problems and their solutions that sensitive attribute data are available as inputs [17, 32, 39, 40, 48, 50, 65].

When labeled data are not available, researchers have made powerful discoveries by augmenting existing data through the inference or construction of labeled data de novo [5, 10, 27, 46]. At times, this work has insufficiently acknowledged the full range of challenges to generating or obtaining those data in applied contexts [35, 76, 77].

Veale and Binns [76] and Kilbertus et al [41] propose approaches to dealing with such information deficits without collecting or revealing sensitive data. Chen et al propose a method to impute unavailable protected class data [13]. Veale et al [77] and Holstein et al [35] outline the contextual needs for implementing fairness. Zliobaite and Custers use theoretical and linear regression-based examples to argue that sensitive data must be included in the modeling process in order to avoid discrimination [79]. Focusing on the European regulatory environment, their study distinguishes between direct and indirect discrimination, but does not address how different sectoral laws enable or prevent detection of either type.

This paper aims to bridge gaps between theoretical approaches and practical constraints, extracting lessons for fair ML practitioners from three real-world case studies.

2 CASE STUDIES

2.1 Credit

United States federal law prohibits creditors from discriminating on the basis of certain protected characteristics. However, across the credit sector, there are sharply divergent approaches to collecting sensitive attribute data. On one hand, mortgage lenders are required to collect such data from their borrowers. On the other, consumer lenders are largely prohibited from doing so. The reason for this difference is not immediately apparent, and likely turns on historical details underlying the development of overlapping legal doctrines.

2.1.1 Background. In the mid-1970s, policymakers acknowledged that discriminatory practices in the consumer credit and home mortgage industries shut out women, people of color, low-income groups, and others from accessing these vital economic resources. The Equal Credit Opportunity Act (ECOA), passed in 1974, initially made discrimination on the bases of marital status and sex illegal, and was later expanded to include other protected groups. The following year, the Home Mortgage Disclosure Act (HMDA) similarly made discriminating against low-income home mortgage borrowers illegal. Like ECOA, HMDA grew to encompass categories including race, gender, and national origin through subsequent amendments.

The origins of ECOA trace back to an era when lenders required unmarried women to have male cosigners on their loans. From the outset, regulators feared that mandatory collection of protected class data beyond gender for the purpose of detecting discriminatory lending might itself facilitate such practices. Thus, under ECOA, the collection of these data is banned. Regulation B, which implements ECOA, has made exceptions for some voluntary collection of data on applicants' color, national origin, religion, race, and sex as "monitoring information" in instances where lenders conduct self-testing to determine whether loans are not being granted to individuals on discriminatory grounds.

The Federal Reserve Board (FRB) twice considered amendments to ECOA that would allow voluntary collection of protected class data for non-mortgage loan applicants in order to surface discriminatory lending decisions. In 1995, the first proposal to lift the ban on collecting sensitive attribute data garnered a mix of support and opposition. Noting that discrimination on protected class bases only covered a limited set of criteria for potential disparate treatment during in-person lending scenarios, supporters of the change pointed to the successful identification and reduction of biased

mortgage lending decisions that resulted from HMDA's strict data collection practices [74]. These advocates disagreed with the FRB's long-held claim that recording these data would lead to discrimination in consumer lending, noting that this predicted harm did not unfold in the home mortgage industry.

The argument that voluntary collection of protected characteristic information would lead to discriminatory lending persisted, however, in large part due to credit industry representatives' complaint letters opposing the amendments. Banks and other lending institutions were not inclined to support a measure that would incur higher costs and stricter reporting standards and presumably may have revealed discriminatory practices. They additionally warned that being asked about sensitive attributes could deter some minority applicants. In response to these public comments following the proposed amendment in 1995, the FRB decided to leave the decision about collecting protected class data up to Congress.

After introducing a second proposal to remove the ban on collecting these data in 1998, the FRB once again determined in 2003 that consumer lending institutions should not gather this information. Standing by their original conviction that sensitive attribute information collection would lead to outright discrimination, the FRB also reasoned that making this a voluntary action could result in incomplete data collection and inconsistent data formatting that would hinder cross-market comparison between creditors [74].

The ECOA's evolution was in many ways the opposite of HMDA's expansive push to seek evidence of unfair practices in the mortgage lending industry. HMDA grew out of home mortgage depository institutions' disproportionate withdrawal of investments in largely urban areas from which they drew their deposits: a form of redlining that devitalized older neighborhoods, since residents could not access the credit required to sell and refurbish their homes [42].

HMDA's initial reporting requirements involved publicizing geographic data about lending patterns. As the contexts and causes of home mortgage lending discrimination changed, HMDA was amended between 1980 and the early '00s to expand the scope of institutions covered and to call for reporting of sensitive attribute data on borrowers' gender, race, income, and other categories. When regulators determined these data were insufficient to demonstrate discrimination, they called for further data collection including data about rejected applications and loan pricing.

2.1.2 Data practices.
While ECOA prohibits collection of sensitive attribute data for most purposes, its implementing regulations allow banks and "anyone who, in the ordinary course of business, regularly participates in decisions about whether or not to extend credit or how much credit to extend" to collect, in a narrow set of circumstances, sensitive attribute data on individuals applying for non-mortgage loans [71]. If lenders opt to collect this data, they must indicate that the information is being recorded for self-testing and monitoring purposes. If an applicant prefers not to provide their race and sex information, the lender is allowed to make their own determinations of these characteristics from visual observation and surname analysis. If the self-test demonstrates that the institution may have violated ECOA, the lender must attempt to identify the cause and extent of the violation. Save for in some instances, the results of the self-test are considered privileged information

that government agencies cannot access in investigations related to ECOA transgressions.

HMDA, by contrast, requires expansive collection of both sensitive attribute data and related mortgage loan application data that can be used to build arguments that discrimination has occurred. Under HMDA, protected data that must be collected as part of a Loan/Application Register (LAR) include sex, race, and ethnicity, with additional requirements that data on income, loan amount and type, property location, and reasons for loan denial (among others) must be reported [42]. Lenders are allowed to use visual observation and surname analysis to guess the sex, race, and ethnicity of applicants who choose not to self-identify these traits. The data are published in different formats depending on the intended recipient. Lenders submit these data to the FRB annually, whereas if a member of the public requested access they would be presented with a modified LAR scrubbed of any identifying information. Finally, the Federal Financial Institutions Examination Council (FFIEC) creates disclosure statements for each lender based on their LAR data, and publishes openly available aggregate reports of HMDA data at city, national, and census-tract levels.

2.1.3 Results and reactions.
The question of whether sensitive attribute data should be collected to detect discrimination in consumer lending remains controversial. As one scholar put it, "Even if computerized credit scoring arguably has the potential to eliminate disparate treatment results, disparate impact discrimination may still occur" [74]. Another scholar has suggested creditors should be *required* to conduct self-testing using sensitive attribute data.[4] Lenders and other proponents of credit scoring systems may argue that expanded collection of data on race and other protected class characteristics would be insufficient to prove discrimination given the increasing complexity of how credit scores are calculated.

Today, as was the case when ECOA was passed, the absence of sensitive attribute data makes it difficult to document and mitigate inequitable consumer lending practices. For example, one of the few, robust public studies on credit scores and discrimination in the United States was performed by the FRB in 2007, at the direction of Congress [56]. To conduct its analysis, the FRB created a database that, for the first time, combined sensitive attribute data collected by the Social Security Administration (SSA) with a large, nationally representative sample of individuals' credit records. The FRB noted its study was unique in part because of the lack of sensitive attribute data in this domain, and this unusual undertaking would not have been possible without significant governmental time and resources.

The shortage of sensitive attribute data in the consumer lending space also complicates regulatory enforcement. For example, in 2013, the Consumer Financial Protection Bureau (CFPB) and the Department of Justice found that Ally Financial, an auto lending firm, overcharged over 230,000 minority borrowers on their car loans. Two years later, the CFPB required Ally Financial to send checks from its $80 million settlement to customers believed to have unfairly paid higher prices for their loans [66]. Lacking access to data on which exact individuals had overpaid, however, the CFPB instead used a Bayesian Improved Surname Geocoding (BISG) method to predict which customers were likely to be racial minorities, and were therefore more likely to be victims of Ally Financial's allegedly

discriminatory pricing. Although BISG's probabilistic means of using publicly available surnames and geographical information as proxies for race and ethnicity is regarded as among the most advanced technique of its kind [11], it is not without flaws. In the use of BISG during the Ally Financial payout, some white Americans were misidentified as having been overcharged for car loans on a discriminatory basis and received compensatory checks [7]. Had data collection practices in non-mortgage lending included sensitive attributes, such mistakes could have been averted. Moreover, predictive power of these techniques might diminish over time if housing and marital segregation patterns change.

By contrast, the amendments to HMDA that spurred collection of protected class data came into effect in 1990, and data from 1992 reflected a significant rise in mortgage lending to low- and moderate-income and minority communities [49]. Moreover, in the longer term, the publication of the 1991 data fueled community activism and helped change home mortgage lenders' practices. Making HMDA data mutually accessible to lending institutions and community organizations is correlated with beneficial outcomes for banks and borrowers alike [20]. However, it remains difficult to know for certain to what extent this data led to reductions in discriminatory lending practices or merely documented changes that were already underway.

2.2 Employment

United States federal law prohibits employers and employment agencies from discriminating on the basis of certain protected attributes. In this context, the collection of demographic information is a familiar part of most employers' day-to-day practice. For example, many large employers are *required* to collect demographic data about job applicants and employees to facilitate regulatory enforcement and research. And for many decades, employment selection procedures have been subject to regulatory guidelines that assume "adverse impact" can be readily quantified.

2.2.1 Background. Following sustained, nationwide demands to end racial discrimination and segregation, Congress passed sweeping protections in the Civil Rights Act in 1964. Title VII of the Act pertains specifically to employment, prohibiting employers from directly or indirectly discriminating in their employment practices and laying out expectations around data collection and reporting for enforcement purposes. The following year, President Lyndon B. Johnson signed Executive Order 11246, which prohibits federal contractors from discriminating in employment decisions, and also requires employers to take affirmative action to increase the representation of women and minorities in their workforces. The order, enforced by the Department of Labor's Office of Federal Contract Compliance (OFCCP), also outlines related requirements around the documentation of recruitment activities, including the collection of demographic information about job applicants and employees in order to facilitate the detection of discrimination at different points in the recruitment pipeline.

Title VII requires employers and other covered entities to "make and keep such records relevant to the determinations of whether unlawful employment practices have been or are being committed," as defined by the Equal Opportunity Employment Commission (EEOC), which enforces the law [1]. Since employers may be liable

for employment practices that result in disparate impact on the basis of protected categories including race and gender, EEOC guidance points to Title VII as a legal basis for requiring the collection of applicant data as necessary to detect, mitigate, or defend against claims of disparate impact. The Uniform Guidelines on Employment Selection Procedures, which reflects the U.S. government's unified position on employment tests, detail how employment tests must be evaluated for unjustified adverse impact on the basis of race, sex, or ethnicity. The EEOC may allow employers to use selection procedures with disparate impact provided that the procedure has been "validated" according to these guidelines [31].

In order to support enforcement of these legal protections, monitor progress in workplace diversity, and enable employer self-assessment, the EEOC also requires private employers with 100 or more employees and contractors with more than 50 employees to collect aggregate statistics about the demographics of their workforce and report them to regulators on a yearly basis, known as EEO-1 reports.

2.2.2 Data practices. Collection of sensitive attribute data in the employment sector is highly standardized, reflecting well-defined federal reporting requirements.

For EEO-1 reports, employers must collect data on sex, a binary field (male or female), as well as race, divided into predefined categories of Black, Hispanic, Asian/Pacific Islander, American Indian/Alaskan Native, white, or "two or more races" [16]. These categories were last updated in 2005 (after 40 years), and in 2007 the EEOC advised that employers were permitted—but not required—to collect more detailed demographic data [55, 63]. Employers must offer employees the opportunity to voluntarily self-identify in the predefined categories. If and only if an employee declines to self-identify, the employer may use "employment records or observer identification," elsewhere described as "visual surveys of the workforce" to categorize the worker to complete their reporting requirements [16].

Although not all employers are required to track sensitive attributes from job applicants, many opt to solicit this information at the time of application, and federal contractors are required to do so. Contractors may solicit demographic data from applicants at any time during the employee selection process so long as the data is solicited from all applicants. Regulators advise that "voluntary self-reporting or self-identification is still generally the preferred method for collecting data on race, ethnicity, and gender, but in situations where self-reporting is not practicable or feasible, observer information may be used to identify race, ethnicity, and gender" [62]. After making "reasonable efforts to identify applicant gender, race, and ethnicity information," contractors may record the applicant's race and gender as "unknown"—with the exception that employers may visually identify applicants "when the applicant appears in person and declines to self-identify" [61]. Notably, employers may not use these data as a part of their employment selection procedures, but may use them to evaluate outcomes and inform changes to those procedures.

2.2.3 Results and reactions. As of 2017, nearly 70 thousand employers file EEO-1 reports per year, documenting data for over 50 million employees [14]. Multiple studies have used EEO and other sources of demographic data to measure trends in occupational

segregation, finding that it has declined since the passage of Title VII [44, 69, 78]. Others use this data to more closely examine race and sex inequality in managerial positions and within specific industries, as well as gender and racial pay gaps [36, 45, 69]. Several researchers were able to determine that OFCCP monitoring and enforcement in particular likely contributed to greater representation of Black workers in skilled occupations [47]. The EEOC and OFCCP themselves commonly use EEO-1 and other mandatorily collected data to support investigations of individual and systemic employment discrimination [15, 75].

Some have pointed out that unlike other government survey instruments, the EEOC merges data on race and ethnicity, which may lead to measurement errors [69]. Others critique the allowance of observed data, but concede that because observed data relate to how workers may be perceived, these data may still have utility in understanding employment discrimination [72]. However, we identified relatively little criticsm of the overall exercise of collecting sensitive attribute data in the context of employment, perhaps because the law requiring and justifying their collection is so clear.

Here again, it is not clear that the relationship between demographic data collection and any occupational desegregation is a causal one. Without this disaggregated employment data, however, documentation of these trends would be significantly more difficult. Indeed, researchers have found that while EEO-1 data do have some constraints, they can be a particularly powerful tool to study workplace inequality and segregation, especially as compared with other data sources [69].

2.3 Health

United States federal law prohibits discrimination in the provision of various health care services. For example, those who qualify for federal health insurance programs such as Medicare or Medicaid may not be subjected to discrimination based on certain sensitive attributes. However, unlike in credit and employment, a major driving factor behind collection of sensitive attribute data in this sector has been voluntary industry efforts to address racial and ethnic disparities in health outcomes, rather than compliance with antidiscrimination laws alone.

2.3.1 Background. The passage of the Civil Rights Act in 1964 and the establishment of Medicare the following year created a need for data to confirm that patients had equal access to health care and that hospitals were not segregated. As a result, many hospitals initially collected data about sensitive attributes for compliance purposes only [68].

A shift in approach was prompted not long after by Secretary of Health and Human Services (HHS) Margaret Heckler's observation in a 1983 national health report that minority health lagged behind that of white Americans, and the subsequent formation of the Task Force on Black and Minority Health to research this gap. The 1986 publication of the Report of the Secretary's Task Force on Black and Minority Health (Heckler Report) marked the first study highlighting the significant health disparities racial minorities experienced in the U.S. [57]. Although the Heckler Report's findings drew awareness to racial inequality in healthcare provision, they

did not themselves effect a shift away from compliance-based sensitive attribute data collection toward a model of using these data to reduce discrimination.

At the request of Congress in 2003, the Institute of Medicine (IOM) published a follow-up report, Unequal Treatment, affirming that unacceptable levels of racial and ethnic disparities in health outcomes persisted [60]. The IOM report concluded that without data on patients' race, ethnicity, socioeconomic status, and primary language, it would be impossible for healthcare providers to detect or address these disparities, and recommended the systematic collection and reporting of race and ethnicity data as a critical step toward eliminating them.

The IOM report jump-started health insurance and other care providers' joint, voluntary effort to collect and use data for healthcare quality improvement and disparity reduction [68]. Organizations such as the National Health Plan Collaborative (NHPC) connected health research institutes to national and regional health plans in order for the former to provide these firms with educational tools and recommendations for how to detect and mitigate discrimination [73]. While initially, many insurance providers believed collecting race and ethnicity data was illegal, legal analysis determined that collection was justified under (though not explicitly required by) Title VI of the Civil Rights Act and the Affordable Care Act, as well as several state laws. Under these statutes, health plans are prohibited from using demographic data for discriminatory purposes, including steering patients toward certain healthcare products [18, 43]. However, health plans are allowed to use these data in order to report aggregate trends and join initiatives to provide equitable services.

2.3.2 Data practices. Some health providers have found it necessary to collect data on patients' race, ethnicity, and primary spoken language (REL) to identify health care disparities [23]. However, there is substantial variability in the precise categories and level of granularity different health providers opt to use to do so. Industry-wide efforts to standardize these data are ongoing.

Physicians and hospitals often collect REL data at intake—usually by asking patients directly, though sometimes determined by intake specialist observation [28]. Health plans, on the other hand, tend to use surveys and incentive programs to collect data after people have signed up for coverage. In some cases, insurers are prohibited from asking for race/ethnicity data during the sign-up process [25, 29]. Some health providers also appear to be able to share and obtain data from federal agencies (e.g. Medicaid), though the exact mechanics of this process remain obscure.

Policymakers and practitioners recognize that in general, data that patients self-report are strongly preferred [24, 33], but in practice, providers have struggled to convince most patients to voluntarily self-report. In the interest of generating data necessary to reduce disparities, methods to estimate race and ethnicity have been widely adopted to supplement self-reported data [54]. Early inference methods involved basic geocoding and surname analysis; more advanced probabilistic techniques have since been developed to refine these estimates. These algorithms produce probabilities that individuals belong to a particular racial or ethnic group, which can then be used to assess disparities between subgroups at an

aggregate level [29, 68]. A number of health plans combine self-reported and estimated data to increase accuracy of their analysis [54].

Experts have recommended that race/ethnicity data based on indirect estimation methods should be stored separately from or be clearly marked in medical systems. Inferred data should not be placed in individuals' clinical medical records—that is, probabilistic methods should not be used to assign someone a particular race or ethnicity classification [68]—but should only be used for aggregate statistical analysis [25]. The IOM recommended that when possible, estimations should be accompanied by their respective probabilities [68]. Whether actual data management practice follows these recommendations likely varies by institution.

2.3.3 Results and reactions. While significant healthcare disparities remain, they have narrowed since the publication of the IOM report that motivated increased data collection [22, 59]. Moreover, granular data has enabled ongoing monitoring and benchmarking of health outcomes, motivated substantial scientific and policy research, and supported federal, state, local, industry, and practitioner-driven disparity reduction initiatives.

Although many health plans have internal policies on confidentiality and use of race/ethnicity data [30], low rates of participation in voluntary data collection may indicate continued lack of trust in healthcare institutions that collect these data, and fear that demographic data might be used to discriminate against patients or otherwise be misused [33]. Health plans have admitted that they sometimes hesitate to collect data for fear of being accused of discrimination [25], on top of other challenges like privacy concerns, IT limitations, and inconsistency or insensitivity in the available categories [54]. But many healthcare providers circumvent these challenges by using techniques to generate demographic data in a probabilistic manner.

Critiques of direct and indirect data collection efforts in healthcare have also emerged on the grounds that concepts of race and ethnicity are merely sociopolitical constructs [25], and therefore categorizing patients using those constructs may reinforce and calcify them. However, the broadly recognized harms of race- and ethnicity-related health disparities seem to have outweighed this critical perspective for the time being.

3 DISCUSSION

Clearly, debates about collection of sensitive attribute data for antidiscrimination purposes are not new. There are decades of precedent that can inform the machine learning fairness research community, the broader technology industry, and other stakeholders.

It is important to reiterate that our case studies do not indicate whether collection of sensitive attribute data has contributed causally to more fair and equitable outcomes. A more fulsome analysis of this question remains for future work. However, we remain convinced that measurement is often a precondition for meaningful improvements.

While the case studies above merely scratch the surface, they offer some important insights. First, they show that U.S. legal frameworks do not offer consistent, extensible guidance about when and how corporations should collect sensitive attribute data. Rather, there are divergent and sometimes contradictory approaches: Some

companies are required to collect sensitive data to comply with antidiscrimination laws, while others are explicitly prohibited from doing so. Second, they show that companies' primary incentives for collecting sensitive attribute data may not—and need not—be compliance or legal requirements at all. The healthcare industry is one such example. Here, deliberate, sustained, and ongoing debates on data collection and inference practices across the industry and stakeholder communities were needed to align on an approach to combating disparities.

If awareness-based techniques remain a primary approach to bias mitigation in predictive modelling, there is a need to thoughtfully consider what efforts must be undertaken to expand collection of sensitive attribute data in a responsible manner.

3.1 Lessons regarding traditionally regulated companies

For traditionally regulated entities like banks and employers, modernization or clarification of laws and regulatory guidance may be needed to encourage the collection of sensitive attribute data for new antidiscrimination efforts. Because these companies can be liable for discriminatory outcomes, they are unlikely to voluntarily collect or analyze sensitive attribute data that could introduce new vectors for liability. Thus, they might resist legal reforms that make it easier to collect sensitive attribute data.

Looking ahead, policymakers, researchers, and and civil society will need to work together to assess what kinds of sensitive attribute data are needed to protect people against discrimination *and* create the policy conditions for that collection to occur. These stakeholders will need to consider what data ought to be collected and in what form, and the appropriate scope of "safe harbor" provisions to incentivize thorough and transparent study. These are not clear or settled questions, even with decades of practice under longstanding civil rights laws.

3.2 Lessons regarding less regulated companies

Many technology platform companies, including those using models to mediate access to important life opportunities, are not squarely covered by civil rights laws. These companies often operate as internet intermediaries, and thus enjoy some special legal protections from liability arising from content posted by third party users [26]. Nonetheless, many are grappling with how to prevent bias. For example:

- Airbnb recently assembled "a permanent team of engineers, data scientists, researchers, and designers whose sole purpose is to advance belonging and inclusion and to root out bias" [52]. The announcement came on the heels of reports of discrimination against African Americans seeking housing opportunities on its platform. The company has not yet publicly discussed the details of this work, or whether it collects or infers sensitive attribute data in its efforts to combat discrimination. However, it is difficult to imagine an approach that would avoid these questions.
- Facebook, in delivering advertisements on its platform, introduces demographic skews along gender and race lines [5]. This practice is currently being challenged in court by the

United States Department of Housing and Urban Development (HUD) [58]. Furthermore, as part of a legally enforceable settlement with civil rights organizations, Facebook recently committed to studying the potential for unintended biases in algorithmic modeling [6]. However, this research will likely be impossible without collecting or inferring sensitive attributes of the company's users. It is not yet clear how Facebook will approach this issue.

- LinkedIn, in an effort to promote equity in hiring, recently updated its recruiter tools to balance the gender distribution in candidate search results, rather than sorting candidates purely by "relevance" [12]. With this update, if the pool of potential candidates who fit an employer's search parameters reflects a certain proportion of women, LinkedIn will re-rank candidates so that every page of search results reflects that proportion. The company also plans to offer employers reports that track the gender breakdown of their candidates across several stages of the recruitment process, as well as comparisons to the gender makeup of peer companies. These features rely on inferring gender data about jobseekers on the platform, which the company was already doing for advertising purposes.

It's not surprising that each of the above examples was motivated by some combination of public pressure or litigation. Technology companies are unsure about what kinds of sensitive attribute collection are appropriate. As a result, the path of least resistance is to simply not to collect or infer data that may create controversy or highlight disparities that may be difficult to address. This is especially true given that perceived violations of privacy are likely to garner intensive media coverage, or where applicable, increased attention from regulators. It will likely fall to a wide range of stakeholders, including advocates, researchers, and policymakers, to ensure that sensitive attribute data is collected and used under appropriate circumstances.

3.3 The need for multidisciplinary collaboration

The implementation of awareness-based antidiscrimination approaches cannot, and should not, move forward without robust involvement of public interest, technical, and regulatory stakeholders. Even amid clear and compelling risks of discrimination or unjust demographic disparities, it can be difficult for policymakers to recommend the collection of sensitive attribute data. There is no evidence this issue will become easier in the future, despite the rapid adoption of machine learning models involved in important life decisions for which these data may be critical to prevent harm.

Privacy laws can sometimes sit in tension with antidiscrimination goals, and might prevent well-meaning actors from collecting data that are necessary to detect and remediate bias in machine learning-based models. Privacy advocates will need to ensure that new legal requirements around data minimization and restrictions on the processing of sensitive data do not deter or impede companies from good faith self-testing and bias remediation. At least one recent U.S. legislative proposal provides an explicit exception for such testing [2], reinforcing the need for more detailed implementation guidance. Meanwhile, European laws and norms diverge

significantly from the U.S. approach, prioritizing privacy heavily over awareness-based antidiscrimination approaches [79]. Private entities may need to navigate conflicting laws, guidance, and public expectations across social and geopolitical contexts.

Finally, there is no shortage of critical questions that still need to be answered:

- *When should sensitive attribute data be collected?* Given the practices described above, non-industry actors should consider under what conditions, if any, they would trust certain private actors with sensitive attribute data that are needed for antidiscrimination efforts. It's obvious that data collection would be justified in some contexts, but the risks may outweigh potential benefits in others. It's far less obvious (and beyond the scope of this paper to suggest) where those lines should be drawn. These norms are especially unsettled for technology companies, who have not had the same historical obligations as traditionally regulated entities, and suffer from significant trust deficits around their data practices.

- *How should sensitive attribute data be created?* Sensitive attribute data can be collected directly from subjects or inferred from non-sensitive data. However, inference presents challenges around consent, forced classification, and error. Stakeholders must work together to determine under what conditions inference is acceptable, appropriate inference methodologies, and how to treat inferred data responsibly. The cases considered in this paper offer instructive approaches, including retaining probabilistic values and uncertainty in inferred data, clearly marking when data are observed or inferred, and storing inferred data separately from data collected with permission. Other approaches might include enforceable commitments to use these data only for detecting and mitigating discrimination.

- *How should sensitive attribute data be treated and secured?* Ideally, sensitive data would be stored separately from other data and used only for limited purposes, but such technical safeguards may be difficult to guarantee. New privacy-protective techniques to access sensitive attribute data, including secure multi-party computation tools like private set intersection and homomorphic encryption, may allow companies to securely sequester these sensitive data from general purpose user data, or even enable trusted third parties to collect, infer, or hold sensitive data while making their insights available to the private entities whose products implicate people's rights [76]. However, these techniques are still nascent and have yet to be broadly deployed for the purpose of bias testing or mitigation.

4 CONCLUSION

Policy debates about the collection and use of sensitive attribute data will decide the fate of awareness-based bias mitigation techniques. There is an urgent need for machine learning scholars to drive these conversations forward, along with other stakeholders, so policy and technical approaches can be developed in accordance with each other. The ability to detect and address bias in algorithms—and the durability of foundational civil rights protections—may hang in the balance.

REFERENCES

[1] [n.d.]. *Title VII, SEC. 2000e-8. [Section 709]*. Retrieved August 21, 2019 from https://www.eeoc.gov/laws/statutes/titlevii.cfm

[2] 2019. Consumer Online Privacy Rights Act (COPRA). https://www.cantwell.senate.gov/imo/media/doc/COPRA%20Bill%20Text.pdf

[3] J. Khadijah Abdurahman. 2019. *FAT* Be Wilin': A Response to Racial Categories of Machine Learning by Sebastian Benthall and Bruce Haynes*. Retrieved August 21, 2019 from https://medium.com/@blacksirenradio/fat-be-wilin-deb56bf92539

[4] Grace Abuhamad. 2019. *The Fallacy of Equating "Blindness" with Fairness: Ensuring Trust in Machine Learning Applications to Consumer Credit*. Master's thesis. MIT Institute for Data, Systems, and Society. https://dspace.mit.edu/handle/1721.1/122094

[5] Muhammad Ali, Aleksandra Korolova, Piotr Sapiezynski, Alan Mislove, Miranda Bogen, and Aaron Rieke. 2019. Discrimination through optimization: How Facebook's ad delivery can lead to skewed outcomes. (2019), 1–15. https://arxiv.org/abs/1904.02095

[6] National Fair Housing Alliance. 2019. *Exhibit A: Programmatic Relief*. Retrieved August 21, 2019 from https://nationalfairhousing.org/wp-content/uploads/2019/03/FINAL-Exhibit-A-3-18.pdf

[7] AnnaMaria Andriotis and Rachel Louise Ensign. 2015. *U.S. Government Uses Race Test for USD80 Million in Payments*. Retrieved August 21, 2019 from https://www.wsj.com/articles/u-s-uses-race-test-to-decide-who-to-pay-in-ally-auto-loan-pact-1446111002

[8] Solon Barocas and Andrew D. Selbst. 2016. Big Data's Disparate Impact. *California Law Review* 104 (2016), 671–732. http://dx.doi.org/10.15779/Z38BG31

[9] Google AI Blog. 2018. *The What-If Tool: Code-free probing of machine learning models*. Retrieved August 21, 2019 from https://ai.googleblog.com/2018/09/the-what-if-tool-code-free-probing-of.html

[10] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of Machine Learning Research*, Vol. 81. 1–15. http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

[11] Consumer Financial Protection Bureau. 2014. *Using publicly available information to proxy for unidentified race and ethnicity: A methodology and assessment*. Retrieved August 21, 2019 from https://files.consumerfinance.gov/f/201409_cfpb_report_proxy-methodology.pdf

[12] Rosalie Chan. 2019. *LinkedIn is using AI to make recruiting diverse candidates a no-brainer*. Retrieved August 21, 2019 from https://www.businessinsider.com/linkedin-new-ai-feature-increase-diversity-hiring-2018-10

[13] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. 2019. Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. 339–348. https://doi.org/10.1145/3287560.3287594

[14] U.S. Equal Employment Opportunity Commission. [n.d.]. *Job Patterns For Minorities And Women In Private Industry (EEO-1)*. Retrieved August 21, 2019 from https://www.eeoc.gov/eeoc/statistics/employment/jobpat-eeo1/

[15] U.S. Equal Employment Opportunity Commission. 2016. *Advancing Opportunity - A Review of the Systemic Program of the U.S. Equal Employment Opportunity Commission: Recruitment and Hiring*. Retrieved August 21, 2019 from https://www.eeoc.gov/eeoc/systemic/review/index.cfm#IIIA

[16] U.S. Equal Employment Opportunity Commission. 2018. *Equal Employment Opportunity Standard Form 100, Employer Information Report EEO-1 Instruction Booklet*. Retrieved August 21, 2019 from https://www.eeoc.gov/employers/eeo1survey/2007instructions.cfm

[17] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17)*. 797–806. https://doi.org/10.1145/3097983.3098095

[18] National Research Council. 2004. *Eliminating Health Disparities: Measurement and Data Needs*. Retrieved December 4, 2019 from https://www.nap.edu/catalog/10979/eliminating-health-disparities-measurement-and-data-needs

[19] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness Through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS'12)*. 214–226. https://doi.org/10.1145/2090236.2090255

[20] National Consumer Law Center et al. 2015. *Group Letter to Consumer Financial Protection Bureau Regarding Public Disclosure of New HMDA Data Points*. Retrieved August 21, 2019 from https://www.nclc.org/images/pdf/foreclosure_mortgage/predatory_mortgage_lending/letter-re-hmda-benefits-and-privacy.pdf

[21] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. 259–268. https://doi.org/10.1145/2783258.2783311

[22] National Center for Health Statistics. 2016. Health, United States, 2015: With Special Feature on Racial and Ethnic Health Disparities. (May 2016), iii–449. https://www.ncbi.nlm.nih.gov/books/NBK367640/pdf/Bookshelf_NBK367640.pdf

[23] Agency for Healthcare Research and Quality. 2012. *Race, Ethnicity, and Language Data: Standardization for Health Care Quality Improvement*. Retrieved August 21, 2019 from https://www.ahrq.gov/research/findings/final-reports/iomracereport/reldata1a.html

[24] Agency for Healthcare Research and Quality. 2012. *Race, Ethnicity, and Language Data: Standardization for Health Care Quality Improvement - 1. Introduction*. Retrieved August 21, 2019 from https://www.ahrq.gov/research/findings/final-reports/iomracereport/reldata1.html

[25] Agency for Healthcare Research and Quality. 2012. *Race, Ethnicity, and Language Data: Standardization for Health Care Quality Improvement - Summary*. Retrieved August 21, 2019 from https://www.ahrq.gov/research/findings/final-reports/iomracereport/reldatasum.html

[26] 47 U.S. Code Section 230. Protection for private blocking and screening of offensive material. [n.d.]. *Section 230. Protection for private blocking and screening of offensive material*. Retrieved August 21, 2019 from https://www.law.cornell.edu/uscode/text/47/230

[27] James Foulds and Shimei Pan. 2018. An Intersectional Definition of Fairness. (2018), 1–16. https://arxiv.org/abs/1807.08362

[28] Robert Wood Johnson Foundation. 2011. *Moving Toward Racial and Ethnic Equity in Health Care*. Retrieved August 21, 2019 from https://www.rwjf.org/en/library/research/2011/03/moving-toward-racial-and-ethnic-equity-in-health-care.html

[29] Allen Fremont, Joel S. Weissman, Emily Hoch, and Marc N. Elliott. [n.d.]. *When Race/Ethnicity Data Are Lacking: Using Advanced Indirect Estimation Methods to Measure Disparities*. Retrieved August 21, 2019 from https://www.rand.org/pubs/periodicals/health-quarterly/issues/v6/n1/16.html

[30] Julie Gazmararian, Rita Carreon, Nicole Olson, and Barbara Lardy. 2012. Exploring Health Plan Perspectives in Collecting and Using Data on Race, Ethnicity, and Language. *The American Journal of Managed Care* 18, 7 (July 2012), e254–e261. https://www.ncbi.nlm.nih.gov/pubmed/22823554

[31] Biddle Consulting Group. [n.d.]. *Uniform Guidelines on Employee Selection Procedures, Section 4: Information on impact*. Retrieved August 21, 2019 from http://uniformguidelines.com/uniformguidelines.html#15

[32] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS '16)*. 3323–3331. https://dl.acm.org/citation.cfm?id=3157469

[33] Romana Hasnain-Wynia and David W Baker. 2006. Obtaining Data on Patient Race, Ethnicity, and Primary Language in Health Care Organizations: Current Challenges and Proposed Solutions. *Health Services Research* 41 (Aug. 2006), 1501–1518. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1797091/

[34] Anna Lauren Hoffman. 2018. Data, technology, and gender: Thinking about (and from) trans lives (as Incoll). In *Spaces for the Future: A Companion to Philosophy of Technology*. Routledge, New York, NY, USA.

[35] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daume III, Miro Dudik, and Hanna Wallach. 2019. Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. 1–16. https://doi.org/10.1145/3290605.3300830

[36] Matt L. Huffman, Philip N. Cohen, and Jessica Pearlman. 2010. Engendering Change: Organizational Dynamics and Workplace Gender Desegregation, 1975–2005. *Administrative Science Quarterly* 55, 2 (June 2010), 255–277. https://doi.org/10.2189/asqu.2010.55.2.255

[37] Ben Hutchinson and Margaret Mitchell. 2019. 50 Years of Test (Un)fairness: Lessons for Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. 49–58. https://doi.org/10.1145/3287560.3287600

[38] IBM. 2019. *AI Fairness 360 toolkit*. Retrieved August 21, 2019 from https://github.com/IBM/AIF360

[39] Faisal Kamiran and Toon Calders. 2011. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33 (2011), 1–33. https://doi.org/10.1007/s10115-011-0463-8

[40] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In *Proceedings of the 35th International Conference on Machine Learning*. 1–5. https://arxiv.org/abs/1711.05144

[41] Niki Kilbertus, Adria Gascon, Matt Kusner, Michael Veale, Krishna Gummadi, and Adrian Weller. 2018. Blind Justice: Fairness with Encrypted Sensitive Attributes. In *Proceedings of the 35th International Conference on Machine Learning*, Vol. 80. 2630–2639. https://arxiv.org/abs/1806.03281

[42] Joseph M. Kolar and Jonathan D. Jerison. 2006. *The Home Mortgage Disclosure Act: Its History, Evolution, and Limitations*. Retrieved August 21, 2019 from https://buckleyfirm.com/uploads/36/doc/HistoryofHMDAapr06.pdf

[43] Sarah Kornblet, Joy Prittsa, Melissa Goldstein, Tom Perez, and Sara Rosenbaum. [n.d.]. Patient Race And Ethnicity Data And Quality Reporting: A Legal. ([n. d.]).

[44] Fidan Ana Kurtulus. 2012. Affirmative Action and the Occupational Advancement of Minorities and Women During 1973-2003. *Industrial Relations* (April 2012), 213–246. https://doi.org/10.1111/j.1468-232X.2012.00675.x

[45] Fidan Ana Kurtulus and Donald Tomaskovic-Devey. 2012. Do Female Top Managers Help Women to Advance? A Panel Study Using EEO-1 Records. *Annals of American Academy of Political and Social Science* 639 (Jan. 2012), 173–197. https://ssrn.com/abstract=1999538

[46] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. *How We Analyzed the COMPAS Recidivism Algorithm.* Retrieved August 21, 2019 from https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[47] Jonathan S. Leonard. 1984. Employment and Occupational Advance Under Affirmative Action. *The Review of Economics and Statistics* 66, 3 (Aug. 1984), 377–385. https://www.jstor.org/stable/1924993

[48] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2019. Fairness through Causal Awareness: Learning Causal Latent-Variable Models for Biased Data. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19).* 349–358. https://doi.org/10.1145/3287560.3287564

[49] Richard D. Marsico. 1999-2000. Shedding Some Light on Lending: The Effect of Expanded Disclosure Laws on Home Mortgage Marketing, Lending and Discrimination in the New York Metropolitan Area. *Fordham Urban Law Journal* 481 (1999-2000), 484. https://pdfs.semanticscholar.org/a009/82327779e40af15816c6cb87d027af40e593.pdf

[50] Aditya Krishna Menon and Robert C. Williamson. 2018. The cost of fairness in binary classification. In *Proceedings of Machine Learning Research,* Vol. 81. 1–12. https://arxiv.org/pdf/1705.09055.pdf

[51] Emanuel D. Moss. 2019. Translation Tutorial: Toward a Theory of Race for Fairness in Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19).* 1–2. https://drive.google.com/file/d/1jKSbnAg7NPotjhvNb8o4YXyjcjF8RTJW/view

[52] Laura Murphy. 2016. *Airbnb's Work to Fight Discrimination and Build Inclusion: A Report Submitted to Airbnb.* Retrieved August 21, 2019 from https://blog.atairbnb.com/wp-content/uploads/2016/09/REPORT_Airbnbs-Work-to-Fight-Discrimination-and-Build-Inclusion.pdf

[53] Arvind Narayanan. 2008. 21 Fairness Definitions and Their Politics. Tutorial presented at the Conf. on Fairness, Accountability, and Transparency.

[54] David R. Nerenz, Rita Carreon, and German Veselovskiy. 2013. Race, ethnicity, and language data collection by health plans: findings from 2010 AHIPF-RWJF survey. *Journal of Health Care for the Poor and Underserved* 24, 4 (Nov. 2013), 1769–1783. https://muse.jhu.edu/article/524354/pdf

[55] Jeffrey A. Norris. 2007. *EEOC Revises Its EEO-1.*

[56] Board of Governors of the Federal Reserve System. 2007. *Report to the Congress on Credit Scoring and Its Effects on the Availability and Affordability of Credit.* Retrieved August 21, 2019 from https://www.federalreserve.gov/boarddocs/rptcongress/creditscore/creditscore.pdf

[57] U.S. Department of Health and Human Services. [n.d.]. *Report of the Secretary's Task Force on Black and Minority Health.* Retrieved August 21, 2019 from https://minorityhealth.hhs.gov/heckler30/

[58] Department of Housing and Urban Development. 2018. *Department of Housing and Urban Development Charge of Discrimination Against Facebook, Inc.* Retrieved August 21, 2019 from https://www.hud.gov/sites/dfiles/Main/documents/HUD_v_Facebook.pdf

[59] Institute of Medicine. 2012. *How far have we come in reducing health disparities?: Progress since 2000: Workshop summary.* The National Academies Press, Washington, DC.

[60] Institute of Medicine Committee on Understanding, Eliminating Racial, and Ethnic Disparities in Health Care. [n.d.]. *Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care.* Retrieved August 21, 2019 from https://www.ncbi.nlm.nih.gov/pubmed/25032386

[61] Federal Contract Compliance Programs Office. 2005. *Obligation To Solicit Race and Gender Data for Agency Enforcement Purposes.* Retrieved August 21, 2019 from https://www.federalregister.gov/documents/2005/10/07/05-20176/obligation-to-solicit-race-and-gender-data-for-agency-enforcement-purposes

[62] U.S. Department of Labor Office of Federal Contract Compliance Programs. [n.d.]. *Internet Applicant Recordkeeping Rule.* Retrieved August 21, 2019 from https://www.dol.gov/ofccp/regs/compliance/faqs/iappfaqs.htm#Q7GI

[63] U.S. Department of Labor Office of Federal Contract Compliance Programs. 2008. *OFCCP Directive No. 283.* Retrieved August 21, 2019 from https://www.dol.gov/ofccp/regs/compliance/directives/dir283.htm

[64] Samir Passi and Solon Barocas. 2019. Problem Formulation and Fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19).* 1–10. https://doi.org/10.1145/3287560.3287567

[65] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '08).* 560–568. https://doi.org/10.1145/1401890.1401959

[66] Corinne Purtill. 2015. *The US is using a race-predicting algorithm to identify victims of car loan discrimination.* Retrieved August 21, 2019 from https://qz.com/537816/in-an-auto-loan-discrimination-case-a-race-predicting-algorithm-is-the-us-governments-best-shot-at-paying-victims-back/

[67] Pymetrics. 2019. *audit-AI (Python library).* Retrieved August 21, 2019 from https://github.com/pymetrics/audit-ai

[68] Agency For Healthcare Research and Quality. [n.d.]. *Race, Ethnicity, and Language Data: Standardization for Health Care Quality Improvement - 3. Defining Categorization Needs for Race and Ethnicity Data.* Retrieved August 21, 2019 from https://www.ahrq.gov/research/findings/final-reports/iomracereport/reldata3.html

[69] Corre L. Robinson, Tiffany Taylor, Donald Tomaskovic-Devey, Catherine Zimmer, and Jr. Matthew W. Irvin. 2005. Studying Race or Ethnic and Sex Segregation at the Establishment Level: Methodological Issues and Substantive Opportunities Using EEO-1 Reports. *Work and Occupations* 32, 1 (Feb. 2005), 5–38. https://doi.org/10.1177/0730888404272008

[70] Andrew D. Selbst, danah boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19).* 59–68. https://doi.org/10.1145/3287560.3287598

[71] Lynn Severson-Meyer. 2003. *Recent ECOA amendments allow lenders an exception for self-testing.* Retrieved August 21, 2019 from https://www.minneapolisfed.org/publications/community-dividend/recent-ecoa-amendments-allow-lenders-an-exception-for-selftesting

[72] Patrick Simon. 2005. The measurement of racial discrimination: the policy use of statistics. *International Social Science Journal* 57, 183 (March 2005), 9–25. https://doi.org/10.1111/j.0020-8701.2005.00528.x

[73] Erin Fries Taylor and Marsha Gold. [n.d.]. *The National Health Plan Collaborative: Overview of Its Origins, Accomplishments, and Lessons Learned.* Retrieved August 21, 2019 from https://www.ahrq.gov/sites/default/files/publications/files/nhpceval.pdf

[74] Winnie Taylor. 2011-12. Proving Racial Discrimination and Monitoring Fair Lending Compliance: The Missing Data Problem in Nonmortgage Credit. *Review of Banking and Financial Law* 31 (2011-12), 218. https://heinonline.org/HOL/P?h=hein.journals/annrbfl31&i=208

[75] Office of Federal Contract Compliance Programs U.S. Department of Labor. [n.d.]. *OFCCP By The Numbers.* Retrieved August 21, 2019 from https://www.dol.gov/ofccp/BTN/index.html

[76] Michael Veale and Reuben Binns. 2017. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data and Society* (July 2017), 1–17. https://doi.org/10.1177/2053951717743530

[77] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18).* 1–14. https://doi.org/10.1145/3173574.3174014

[78] Kim A. Weeden, Mary Newhart, and Dafna Gelbgiser. 2018. State of the Union 2018: Occupational Segregation. *The Poverty and Inequality Report (Stanford Center on Poverty and Inequality)* (2018), 30–33. https://inequality.stanford.edu/sites/default/files/Pathways_SOTU_2018_occupational-segregation.pdf

[79] Indre Zliobaite and Bart Custers. 2016. Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. *Artificial Intelligence and Law* 24, 2 (June 2016), 183–201. https://link.springer.com/article/10.1007/s10506-016-9182-5