Demographic-Reliant Algorithmic Fairness: Characterizing the Risks of Demographic Data Collection in the Pursuit of Fairness

McKane Andrus*
Sarah Villeneuve*
mckane@partnershiponai.org
sarah.v@partnershiponai.org
Partnership on AI
San Francisco, California, USA

ABSTRACT

Most proposed algorithmic fairness techniques require access to demographic data in order to make performance comparisons and standardizations across groups, however this data is largely unavailable in practice, hindering the widespread adoption of algorithmic fairness. Through this paper, we consider calls to collect more data on demographics to enable algorithmic fairness and challenge the notion that discrimination can be overcome with smart enough technical methods and sufficient data. We show how these techniques largely ignore broader questions of data governance and systemic oppression when categorizing individuals for the purpose of fairer algorithmic processing. In this work, we explore under what conditions demographic data should be collected and used to enable algorithmic fairness methods by characterizing a range of social risks to individuals and communities. For the risks to individuals we consider the unique privacy risks of sensitive attributes, the possible harms of miscategorization and misrepresentation, and the use of sensitive data beyond data subjects' expectations. Looking more broadly, the risks to entire groups and communities include the expansion of surveillance infrastructure in the name of fairness, misrepresenting and mischaracterizing what it means to be part of a demographic group, and ceding the ability to define what constitutes biased or unfair treatment. We argue that, by confronting these questions before and during the collection of demographic data, algorithmic fairness methods are more likely to actually mitigate harmful treatment disparities without reinforcing systems of oppression. Towards this end, we assess privacy-focused methods of data collection and use and participatory data governance structures as proposals for more responsibly collecting demographic data.

CCS CONCEPTS

• **Security and privacy → Social aspects of security and privacy**; *Privacy protections*; *Economics of security and privacy*; • **Social and professional topics → User characteristics**; *Gender*; *Sexual orientation.*

KEYWORDS

demographic data, sensitive data, categorization, fairness, discrimination, identity, race, gender, sexuality, measurement

ACM Reference Format:
McKane Andrus and Sarah Villeneuve. 2022. Demographic-Reliant Algorithmic Fairness: Characterizing the Risks of Demographic Data Collection in the Pursuit of Fairness. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3531146.3533226

1 INTRODUCTION

As organizations increasingly adopt algorithmic decision-making systems (ADMS) for their efficiency and purported objectivity, harmful and discriminatory decision patterns have been observed in contexts such as healthcare [80], hiring [26, 91, 104], criminal justice [121], and education [96]. In response, a swath of algorithmic fairness strategies have been proposed to better understand how ADMS treat certain individuals and groups in order to detect and mitigate harmful biases.

Most current algorithmic fairness techniques require access to demographic data (such as race, gender, or sexuality) in order to make performance comparisons and standardizations across groups [116]. These demographic-based algorithmic fairness techniques look to overcome discrimination and social inequality with novel metrics operationalizing notions of fairness and by collecting the requisite data, often removing broader questions of governance and politics from the equation [5]. This paper problematizes this approach, arguing instead that collecting more data in support of fairness is not always the answer and can actually exacerbate or introduce harm for marginalized individuals and groups. We believe more discussion is needed in the machine learning community around the consequences of "fairer" algorithmic decision-making, and what conditions must be satisfied in order responsibly collect and use demographic data for fairness purposes.

The rest of the paper proceeds as follows: In Section 2 we define what we mean by demographic data. Sections 3 and 4 outline the importance of demographic data for addressing algorithmic discrimination and the challenges practitioners currently face when attempting to collect and use this data for fairness purposes. In Sections 5 and 6 we characterize the risks that can emerge through the collection and use of demographic data to individuals and to communities at large. Finally, in Section 7 we outline promising

*Both authors contributed equally to this research.

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06…$15.00
https://doi.org/10.1145/3531146.3533226

approaches to mitigating a number of the risks discussed in this paper.

2 DEFINING DEMOGRAPHIC DATA

Throughout the literature on algorithmic fairness, we see a range of terms used to refer to data about demographics, such as *demographic attributes*, *protected categories/classes/characteristics*, and *sensitive features*, among others. In most cases, these terms are used to refer simply to the variables in a dataset for which different fairness metrics should be computed, without much nuance regarding what is actually captured by the variable. Concepts such as gender, race, ethnicity, and sexuality are collapsed into single categories that are seen as self-evident, inherent characteristics of one's identity. This approach is understandable, seeing as most of the points of reference regarding anti-discrimination, or fairness more broadly, in machine learning stem from legal notions of discrimination that prescribe standards of fairness using census-like data [12]. Single variable representations of gender, race, ethnicity, and other demographic categories are schemas of social categorization that are both highly measurable and highly actionable, as they tell a relatively clear story about what types of differences exists across groups. What they often miss, however, is how these schemas are themselves products of social systems and conceal parts of the story when taken at face value – to be Black in the United States does not just mean having the "attribute" of dark skin, it also means being on the receiving end of generations of oppression and disenfranchisement through social, governmental, and political means.

If not just a demographic variable in a dataset and not just an individual characteristic, how then should we think about the concepts underlying demographic data like race or gender? In this work, we draw from long histories of scholarship that interrogate categorization schemas and the social harms that can stem from their uncritical adoption [17, 32, 44, 45, 48, 54, 60, 126]. *Stemming from these understandings, we see demographic data, and demographic categories more broadly, as an attempt to collapse complex social concepts into categorical variables based on observable or self-identifiable characteristics.* While data of this kind can certainly help establish claims of unfairness, as we explore in many of the sections below, the mismatch between efforts to make these categories legible to computers and the actual, multidimensonal, and often fluid nature of class-membership can undermine work around fairness from the start.

3 ALGORITHMIC FAIRNESS, DISAGGREGATED EVALUATIONS, AND DEMOGRAPHIC DATA

As algorithmic decision-making systems become more widespread, there is greater risk for the systems to reinforce historical inequalities and engender new forms of discrimination in ways that are difficult to assess. In many cases, when ADMS discriminate against protected groups, they do so indirectly [125]. While it is certainly possible for machine learning systems to base decisions off of features like race, more often the tools uncover trends and correlations that effectively discriminate across groups without relying on demographic variables.

When looking at how algorithms can discriminate, it is important to consider the different ways in which bias can enter the picture. The most often discussed point of entry is the data used to build the system. Biases in the data collection process and existing social inequalities will dictate the types of correlation that can be utilized by a machine learning system. If a group is underrepresented in the dataset or if the dataset embeds the results of historical discrimination and oppression in the form of biased features, it is to be expected that ADMS will have worse performance for or undervalue certain groups [78, 83].

How ADMS are designed and towards what kinds of objectives, on the other hand, can have a large bearing on how discriminatory their outcomes are [86]. If optimizing for a goal that is poorly defined, or even discriminatorily defined, it is likely that a system will reproduce historical inequity and discrimination, just under a guise of objectivity and disinterestedness [103]. For example, the UK higher education admission algorithm that attempted to define aptitude as a combination of a predicted performance and secondary school quality systematically biased the outcomes for those coming from poorer or less-established secondary schools [96]. Similarly, ADMS that ignore contextual differences between groups in an attempt to treat everyone equally often lead to discriminatory outcomes, such as in the case of hate speech detection systems that do not consider the identities of the speaker [30, 31].

Though the types of discrimination discussed here represent a small subset of the myriad ways that ADMS can discriminate, they still bring up a difficult question — how should practitioners assess the potential discriminatory impacts of their systems? The nascent field of Algorithmic Fairness has contributed a number of strategies for identifying and even mitigating discrimination by ADMS, however, almost all of the proposed methods require the use of datasets which include the potentially discriminated against demographic attributes. Even outside of Algorithmic Fairness, conducting any kind of quantitative evaluation that disaggregates results across groups is likely to require data on group membership [10].

As previous research has highlighted, data on demographic categories is often unavailable due to a range of organizational challenges, legal barriers, and practical concerns [3]. Some privacy laws, such as the EU's General Data Protection Regulation (GDPR), effectively only allow the collection of sensitive data such as race, religion, and sexuality under strict conditions of meaningful consent from data subjects [41, 112]. Some corporate privacy policies and standards, such as Privacy By Design, call for organizations to be intentional with their data collection practices, only collecting data they require and can specify a use for [3]. Given the uncertainty around whether or not it is acceptable to ask users and customers for their sensitive demographic information, most legal and policy teams urge their corporations to err on the side of caution and not collect these types of data unless legally required to do so. As a result, concerns over privacy often take precedence over ensuring product fairness since the trade-offs between mitigating bias and ensuring individual or group privacy are unclear [3].

Furthermore, prior work has shown that demographic data is generally only collected once a narrow, enforceable definition of discrimination is codified into law or corporate standards [16]. As

such, the issue of missing demographic data is often only confronted and explicitly addressed once assessment and/or enforcement efforts begin in earnest[1]. Even then, however, we see that anti-discrimination standards and practices vary widely across domains, and in many cases specific types of discrimination are legally sanctioned (e.g., "actuarial fairness" in insurance quotes and "legitimate aims" in employment law [81]). Most often, however, legal anti-discrimination frameworks consider ignoring or omitting demographic attributes altogether non-discriminatory [124]. When ADMS use this approach, often called "fairness through unawareness" or (in cases involving race) "color-blindness," the results have often been shown to be just as discriminatory as whatever came before algorithmic decision-making [65]. In other cases still, antidiscrimination law and policies can indirectly inhibit corporations from using demographic data, even if it is permitted, since being made aware of discrimination opens the door to legal liability if the discrimination is uncovered without a plan to successfully mitigate it [3].

Beyond uncovering bias and discrimination, access to demographic data can help provide justification for the adequate representation and participation of various groups during the design and implementation of ADMS. The trajectory of COVID-19 data collection in the U.S. serves as a good example of this – though the CDC requested racial demographic data to be collected on everyone who was treated for symptoms of COVID-19, racial demographics were frequently omitted in most local and state data collection efforts [6]. As such, the unique vulnerabilities of Black, Indigenous, and Latinx individuals and communities against the virus were largely obscured until data collection and inference methods improved [6].

Partially as a result of this general absence of demographic data, we frequently see a cycle of ADMS development and deployment, exposure of egregious discrimination through individual reports, and then ad hoc system redesigns. [2] Without access to demographic data, it is difficult to assess these types of shortcomings before system deployment, and even after deployment it is likely that more insidious forms of discrimination remain hidden.

In the face of such a cycle, organizations looking to assess algorithmic fairness techniques have called for guidance on how to responsibly collect and use demographic data. However, prescribing adherence to statistical definitions of fairness on algorithmic systems without accounting for the social, economic, and political systems in which they are embedded can fail to benefit marginalized groups and undermine fairness efforts [4, 74]. Therefore, developing guidance requires a deeper understanding of the risks and tradeoffs inherent to the use of demographic data. Efforts to detect and mitigate harms must account for the wider contexts and power structures that algorithmic systems, and the data that they draw on, are embedded in.

[1]See, for example, the new data collection efforts mandated by Exec. Order No. 13985 and Exec. Order No. 14035

[2]The AI Incident Database [85] includes many examples of this, such as Google just removing the 'gorilla' tag after it was applied to Black users' photos [49] and Amazon scrapping their resume screening tool after it was shown to penalize experience with organizations such as "Women Who Code" [43].

4 COMMON CONCERNS AROUND MEASURING DEMOGRAPHICS

When considering the risks of demographic data collection and use, it is often important to consider how the data is collected. Generally speaking, some combination of self-identification, ascription, and inference are used to create datasets that include demographics or to supplement existing datasets with demographic categories. Self-identification is usually operationalized as self-classification, or having data subjects select relevant categories from a set of options, ascription, also referred to as labelling, relies on data labelers or other second parties to determine the data subjects' demographics from existing data sources such as images or text, and inference or imputation use statistical methods and machine learning to guess subjects' attributes based on correlations found in datasets that already included demographic variables. Each of these approaches comes with trade-offs around privacy, data quality, and technical and economic feasibility that must be balanced when deciding whether to, and if so how to, collect demographic data [3].

To start, each of these approaches takes on a unique set of risks to personal privacy. Self-identification arguably incurs the least risk, as requiring data subject participation ensures that they have more control over and awareness of what types of data about them exists. That being said, it can also have the impact of making people more aware of their privacy and increase concerns about what the data will be used for, an outcome many corporations try to avoid. Ascription and inference, on the other hand, allow data holders to define aspects about data subjects without giving them a say or even without them being made aware. These methods can differ in risk in that some individuals might be more comfortable having attributes like their gender identity predicted by an algorithm instead of say a data labeler given the growing disillusionment and dismissal of algorithmic profiling [21]. Furthermore, while ascription and inference can perhaps mitigate contests over privacy on the front end, they greatly increase the risk of public backlash if or when it is revealed that this data was collected, such as in the case of Facebook inferring "ethnic affinity" or "multicultural affinity" [57].

Each measurement technique also comes with unique challenges around the quality of the resulting datasets. While self-reporting likely results in the most accurate labels, it can also produce the sparsest datasets, as many individuals will not share sensitive attributes unless they are sufficiently incentivized or bought into the goal of collection [3, 63]. In cases where ascription is used, on the other hand, datasets are more complete but much less accurate. Depending on the types of data available when ascribing attributes, whether it be images, written text, or some other source of metadata about a person, certain demographic categories may just not be at all determinable (e.g. sexuality when the available data is just images of faces). Inference techniques encounter a similar tradeoff of completeness for accuracy, but they are more scalable and so find more use in practice. Inference is also commonly used as a supplemental technique, allowing practitioners to fill in demographic attribute variables in incomplete datasets. While this is a low-cost strategy that can enable many kinds of algorithmic fairness analysis, it requires practitioners to be mindful of how inferred attributes introduce new sources of bias to the analysis [25].

Though concerns around privacy and data quality are deeply pervasive, costs and organizational risks are the most likely barriers to collecting demographic data [3]. In the case of self-identification, for many companies asking users or clients for their demographics would be seen as questionable if not outright nefarious. As such, the effort needed to communicate how demographic data would be used as well as the potential legal risks from privacy and anti-discrimination law simply set too high of a cost on collecting demographics through self-identification. Ascription, on the other hand, is generally cost-prohibitive for large datasets because it relies on paid employees and contractors to produce demographic information. That being said, it can be used to create small, high quality datasets for assessing system bias in cases where discrimination is likely to be based on ascribed characteristics, such as perceived gender or race [13]. Inferential methods, despite carrying the most risk of inaccuracy when used by themselves, are extremely common in practice largely because of their immediate feasibility. In many domains it is only explicitly asking individuals for their sensitive attributes that is legally prohibited, not the actual use of those attributes to assess discrimination. As such, practitioners looking to assess their systems for discriminatory behavior are pushed to create what are known as *proxy models* that predict sensitive attributes to assess potential discrimination [20, 24, 58]. Without explicit guidance or requirements around algorithmic antidiscrimination, this type of measurement strategy is likely to become more pervasive despite incurring some of the greatest risks to accuracy and privacy.

5 INDIVIDUAL RISKS OF OF DEMOGRAPHIC DATA COLLECTION AND USE

When discussing the social risks of collecting demographic data, most researchers and practitioners focus mainly on the threats to individual privacy. In this section we expand on the individual privacy conversation and consider two more sources of risk – individual misrepresentation and use beyond intended consent. Our goal with this section and the next is not to suggest that demographic data should never be used, but rather to build out a clearer picture of what future data collection efforts should attempt to address in their pursuit to enable less discriminatory decision making.

5.1 Encroachments on Individual Privacy

Though privacy is a commonly held concern when it comes to any type of data collection, the collection of demographic data requires special care and consideration. Sharing or otherwise determining sensitive attributes can expose individuals to various forms of direct or indirect harm, especially already marginalized and vulnerable individuals. Though there are numerous proposed methods for ensuring the privacy and security of sensitive attributes, the strategies for assessing (let alone mitigating) fairness or discrimination under privacy constraints are still very experimental and not commonly used [37, 56, 64]. As such, most efforts to collect sensitive demographic data will at some point in the pipeline require tying the data to individuals, necessarily risking individuals' privacy.

One clear privacy risk of obtaining an individual's demographic data is that this data can still be the basis for many types of discrimination. Though many countries have laws against direct discrimination, it is still a common occurrence due to the difficulty of proving discrimination in individual cases, especially algorithmic ones [118]. In domains such as hiring [89, 90], advertising [22, 29], and pricing [55, 72], direct forms of discrimination, algorithmically mediated or not, are relatively common. For sectors like advertising, discriminatory practices are often justified by claims that differential treatment results in more helpful services, which may in fact be true. However, in a recent survey study of Facebook users, most were still uncomfortable with sensitive attributes being used as the basis for decisions around what they are being shown [22].

In the most pernicious cases, demographic data can be used as the criteria for various forms of state or societally enacted violence, such as detainment and deportation based on documentation status in the United States. Even in cases where the targeted attribute (e.g., documentation status) is not collected, other accessible forms of data (e.g., country of birth and spoken language, contact lists) can be used to help infer the targeted attribute [111]. As corporately collected data becomes increasingly requested by and made available to state agencies [66, 99, 109], it is critical that practitioners consider what types of identity-based violence individuals might be exposed to by providing self-categorizations.

Depending on what types of threat individuals' feel behind having their sensitive attributes revealed, it is also possible that the collection of demographic data can have a "chilling effect" on members of the groups most at-risk of discrimination or targeting. Once cognizant of the possibility that a platform or system is directly asking for or inferring demographic attributes, individuals may change their behavior on a platform or with a system to prevent being labeled or out of concerns such as "stereotype threat," the concern of being viewed as an example of a negative stereotype about some aspect of one's identity [120]. A commonly suggested approach to reducing these forms of direct targeting risk is to "anonymize" or "de-identify" datasets, but even with these strategies marginalized individuals can still be vulnerable to "re-identification" [23].

Finally, another salient privacy risk is the possible loss of autonomy over one's identity and interactions when demographic data is collected or used. Machine learning and AI systems are often built with the intention of making generalizations across groups in order to categorize individuals, meaning that it is not even necessary for an individual to share their demographic data in order for the system to decide to treat them as a "Black woman" or "Asian man." Simply by matching patterns of behavior, algorithmic systems can categorize individuals, even if the categories are not explicitly labeled "Black woman" or "Asian man" [70, 75]. Barocas and Levy [11] refer to these types of associations between individuals as privacy dependencies, as an individual's privacy quite literally depends on the privacy of the people like them. In other cases, even when users provide sensitive data about themselves, platforms may not take that data into account when making decisions for that user, subverting their agency around self-presentation [15].

For these types of privacy risks, we might expect privacy regulation such as GDPR or California's California Consumer Privacy Act (CCPA) to prevent the worst abuses. Privacy regulation to date, however, has largely focused on the individual's "right to privacy"

and agency over their own personal data [75], ignoring the relationality between data subjects and the interconnectedness of their privacy. Even when it comes to an individuals' agency over data about them specifically, the relationship between individuals and the tech firms collecting their data is frequently one of "convention consent" [110]. In other words, users are resigned to share their data even when they do not agree with how it is being used because it is the cost of accessing platforms and services and they do not see any reasonable alternative [33]. While there is technically always the option of not using platforms or services that require personal data, many have come to serve as essential infrastructure, calling into question how much someone can afford to hold onto their privacy by withholding their consent.

5.2 Miscategorization and Identity Misrepresentation

How demographic data is coded and represented in datasets — specifically, what categories are being used to define individual characteristics — can have a significant impact on the representation of marginalized individuals. When ADMS fail to accurately determine an individual's identity, such miscategorization and identity misrepresentation may not only lead to social and political discrimination, but also psychological and emotional harms via feelings of invalidation and rejection [39].

On one hand, individual miscategorization can occur when an individual is misclassified despite there being a representative category that they could have been classified under. To better understand the implications of miscategorization, it's important to understand the different dimensions of identity and how these can lead to misrepresentation. With respect to racial identity, Roth [98] distinguishes between multiple dimensions of the concept of race, highlighting how an individual's racial identity can be represented differently depending on the observer or method of data collection. Dimensions of racial identity include self-identity (the race an individual self-identifies as), self-classification (the racial category an individual identifies with on an official form), observed race (the race others believe you to be), appearance-based (observed race based on readily observable characteristics), interaction-based (observed race based on characteristics revealed through interaction such as language, accent, surname), reflected race (the race you believe others assume you to be), and phenotype (racial appearance) [48]. When an individual is categorized under just one of these dimensions, unless use of the data is limited to a single bespoke purpose, it is highly likely that the individual will be misrepresented during disaggregated analyses in some way. For example, when one's self-classified race is collected and it differs from their most frequent perceived race, the analysis is likely to miss forms of discrimination stemming from perceived race [67].

On the other hand, identity misrepresentation can occur when the categories used do not adequately represent an individual as they self-identify. As Keyes et al. [61] argues, ADMS designers and the data used carry particular expectations of what gender, class, or race mean in society. When categorization and classification of an individual is conducted by observation, either by person or machine, there is the risk that an individual's observed identity does not align with their self-identification and can lead to individual misrepresentation. Moreover, treating the notion of identity as a quality that can be "inferred" externally produces new forms of control over an individual's agency to define themselves [59–61]. For example, ADMS that involve predicting an individual's sexual identity perpetuate certain beliefs and ideas about queerness by associating specific characteristics, appearance, biology, or behavior as essential features of sexual identity [111]. This can cause psychological harm to individuals who may not "fit the mold" of the category they self-identify with. As individuals come to understand the differences that form the basis for categorization, they can start to interpret their own actions through the lens of the category they are assigned to, in turn influencing their future decisions, a process that philosopher Ian Hacking dubbed the "looping effect" [45]. For example, when individuals are made more acutely aware of what factors lead to them being perceived as "a woman" or as "queer," they are incentivized to change their behavior either to increase the likelihood of their preferred classification or to simply live in a way they may now see as more aligned with their identity [32]. Though this type of risk is not likely to be the most salient when collecting demographic data only to assess unequal outcomes or treatment, it is important to be mindful of when asking users for their demographics on platforms with content recommendations that are increasingly tailored to users based on the other information they provide, such as YouTube and TikTok.

Additionally, restricting identity to fixed and measurable forms inherently misrepresents fluid and often unobserved characteristics such as sexuality and gender identity [100, 111]. Facial recognition technologies are a prominent case where the harm of identity misrepresentation occurs, since categorization is often based solely on observable characteristics. Many datasets used to train facial recognition systems are often built upon a binary, physiological perspective of female and male, and consequently misrepresent individuals who do not self-identify with those categories [102]. Continuing to build databases that assume identity is fixed and that only include observable traits risks reinforcing harmful practices of marginalization. Additionally, doing so can further entrench pseudoscientific practices which assume invisible aspects of one's identity from visible characteristics such as physiognomy [102, 107].

5.3 Data Misuse and Use Beyond Informed Consent

Once collected, demographic data can be susceptible to misuse. Misuse refers to the use of data for a purpose other than that for which it was collected or consent was obtained. In the context of ADMS, this could involve collecting and using data to train models that may be deployed in unexpected contexts or re-purposed for other goals. In practice, it is difficult for organizations to specify clear data uses at the point of collection. Sensitive demographic data, in this case, can go on to inform systems beyond the initial scope defined during collection. For example, in 2019 the U.S. government developed the Prisoner Assessment Tool Targeting Estimated Risk and Needs (PATTERN) [113]. PATTERN was trained on demographic data and criminal history data for the purpose of assessing recidivism risk and providing guidance on recidivism reduction programming and productive activities for incarcerated people [113]. Then, in March 2020 the Bureau of Prisons was directed to begin

using PATTERN to determine which individuals to transfer from federal prison to home confinement in the wake of the COVID-19 pandemic [84]. However, the data used to inform PATTERN was not intended to inform inmate transfers, let alone during a global pandemic which introduced a number of unprecedented social and economic variables [84].

Data misuse could also refer to instances where data is shared with third parties or packaged and sold to other organizations. A notable example of data misuse in this respect can be seen in Clearview AI's facial recognition dataset, which the company claims contains over three billion images scraped from social media platforms such as Facebook, Instagram, LinkedIn, and Twitter, along with personal information listed on people's social media profiles [50].

Corporations collecting and using individuals' demographic data to train and deploy ADMS are facing some increased pressure (from both the public and regulatory bodies) for transparency on how such data is collected and used. For example, Article 13 of the GDPR [36] requires companies collecting personal data from a data subject to provide the data subject with information such as the purpose of the data processing, where the data is being processed and by which entity, recipients of the data, the period for which the data will be stored, the existence of algorithmic decision-making and the logic involved, and the right to withdraw data [36]. Companies have begun to incorporate this informational requirement into their data collection practices, often in the form of click wraps, digital banners that appear on users' screens and require them to "accept all" or "decline" a company's digital policies. Yet, providing individuals with transparency and information about how data will be used is generally not sufficient to ensure adequate privacy and reputational protections [79]. Overloading people with descriptions of how their data is used and shared and by what mechanisms is not a way to meaningfully acquire data subjects' consent, especially in cases where they are sharing sensitive, personal information [79, 82]. Rather, the goals of data use and the network of actors expected to have access to the data are what need to be clearly outlined and agreed upon by the data subject. Additionally, while it may be difficult for organizations to specify clear data uses at the point of collection, companies may consider providing updates as the use cases for that data becomes clearer. In following with this more rigorous notion of consent, we would expect check-ins on how the data was used to assess or mitigate discrimination and on whether the data subjects would still like for their sensitive data to be used towards these ends. Collecting demographic data consensually requires clear, specific, and limited use as well as strong security and protection following collection.

6 COMMUNITY RISKS OF DEMOGRAPHIC DATA COLLECTION AND USE

Moving beyond individual risks, this section considers a range of potential harms to communities. As ADMS seek to generalize across groups based on data collected from a subset of the population, data collection can lead to a number of unintended risks including undue surveillance, group misrepresentation, and the ceding of agency over defining what constitutes fair and just treatment, which we detail below.

6.1 Expanding Surveillance Infrastructure in the Pursuit of Fairness

Data collection is now employed on a regular basis to define and monitor types of groups, such as customers, communities, or populations. This type of surveillance does not target individuals directly, but looks at how people can be grouped together or what it means to be a member of a specific group. This form of group profiling raises a number of questions around harms related to collective privacy and discrimination [70, 75].

As discussed throughout this paper, there is often a trade-off between privacy and fairness when it comes to assessing discrimination and inequality. Calls to collect demographic data in order to enable algorithmic fairness techniques run the risk of intersecting with many corporations' and governments' attempts to understand the impacts their products and services are having on groups at large, potentially justifying the expansion of data-driven surveillance infrastructures. Scholars of surveillance and privacy have shown time and time again that the most disenfranchised and "at-risk" communities are routinely made "hypervisible" by being subjected to invasive, cumbersome, and experimental data collection methods, often under the rationale of improving services and resource allocation [14, 19, 35]. Within this context it is not unreasonable for members of disenfranchised groups to distrust new data collection efforts and to withhold information about themselves when sharing it is optional.

Further still, increased visibility and awareness of being under surveillance is likely to have a chilling effect on community groups and society at large. Citron and Solove [27] highlights that data-based surveillance can reduce the range of viewpoints and amount of information shared among communities. One example of this is the dramatic decrease of Grindr users sharing their HIV status on the app when it was learned that Grindr had shared this data with analytics firms [27]. In this way, attempts to gain insight into specific groups through demographic data collection may result in widespread self-censoring. Though most practitioners are well-meaning in their efforts to improve representation and system performance for groups, it is important to consider what the costs and risks are for already disenfranchized groups to be 'better' represented in datasets [52].

In cases where there seems to be a trade off between institutional visibility or anti-discrimination and surveillance, we recommend centering the agency of the groups that planned interventions are supposed to support. Scholarship from the emerging fields of Indigenous Data Sovereignty and Data Justice can provide a starting point for what this might look like — instead of collecting demographic data to "objectively" or "authoritatively" diagnose a problem in the system or even in society more broadly, data collection efforts can be grounded in community needs and understandings first and foremost [92, 95, 119]. It is also important to note that disaggregated data is not the only way that groups facing discrimination or other forms of inequality can become more visible. Small-scale data collection and qualitative methodologies can also be used to identify treatment and outcome disparities [69, 97, 122].

6.2 Group Misrepresentation and Reinforcing Oppressive or Overly Prescriptive Categories

Another source of risk arises from the demographic categories themselves and what they are taken to represent. Scholars from a wide range of disciplines have considered the question of what constitutes representative or useful categorization schemas for race, gender, sexuality, and other demographics of institutional interest and where there are potential sources for harm [17, 32, 44, 45, 48, 54, 60, 126]. Though there are certainly nuances to defining and measuring each of these demographics, we can find some general trends across this scholarship around the risks of uncritically relying on these categories to describe the world, or, in our case, to ascertain treatment differences across groups. At a high level, these risks center around essentializing or naturalizing schemas of categorization, categorizing without flexibility over space and time, and misrepresenting reality by treating demographic categories as isolated variables instead of "structural, institutional, and relational phenomenon" [48, 1].

The first of these risks, and certainly the one most frequently encountered and vocalized by practitioners [3], is when entire groups are forced into boxes that do not align with or represent their identity and lived experience. Often, this occurs because the range of demographic categories is too narrow, such as leaving out options for "non-binary" or "gender-fluid" in the case of gender [15]. It can also commonly occur in cases where demographic data is collected through inference or ascription. In these cases, systems often embed very narrow standards for what it means to be part of a group, defining elements of identity in a way that does not align with the experience of entire segments of the population [39]. This type of risk is especially well-documented with regards to various types of automated gender recognition failing to correctly categorize transgender and non-binary individuals. Both critics and users deem these failures inevitable because these systems treat gender as purely physiological or visual, which is different from how members of these communities actually experience gender [39, 47, 59, 61]. In each of these ways, demographic data collection efforts can reinforce oppressive norms and the delegitimization of disenfranchised groups, potentially excluding entire communities from services and institutional recognition as a form of what critical trans scholar Dean Spade calls "administrative violence" [106].

Furthermore, data collected with too limited of categories risks misrepresenting and obscuring subgroups subject to distinct forms of discrimination and inequality, especially in cases where demographic data is collected via inference. The most common inference techniques used by public and private institutions generally rely on the very limited set of demographic categories included in the census, such as Bayesian Improved Surname Geocoding (BISG), which uses an individual's name and zip code to predict their race [34]. As one example, there have been many efforts to distinguish between Asian American and Pacific Islander (AAPI) populations in health [105] and education [88] due to fears that disenfranchised subgroups are made further invisible by being categorized under the broad umbrella of AAPI. Models like BISG, however, use U.S. census data and thus cannot go beyond the six census categories for race and ethnicity (White, Black, AAPI, American Indian/Alaskan Native, and Multiracial).

Another way that categorization schema can be misaligned with various groups' experiences and lived realities is when the demographic categories themselves are too narrowly defined to capture all the dimensions of possible inequality. For example, each of the dimensions of race discussed in Section 5.2 carries with it different potential adverse treatments and effects. If the only type of demographic data an institution collects is through self-identification, for instance, it can draw a very different picture of discrimination than data collected through ascription [48, 101]. As such, when it comes to assessing discrimination or some other form of inequality, it is critical that practitioners have a prior understanding of how differential treatment or outcomes are likely to occur such that the right dimension of identity is captured to accurately assess likely inequities.

As previously mentioned, it is also important to consider the temporality of categorization — categorization schema and identities can change over time, and how much this is taken into account during system design will likely have a disproportionate impact on groups with more fluidity in their identities. Looking first to gender and sexuality, critical data scholars have argued that queer and trans identities are inherently fluid, contextual, and reliant upon individual autonomy [60, 100]. There are no tests or immutable standards for what it means to be queer, non-binary, or any number of other forms of identity, and it is likely that one's presentation will change over time given new experiences and contexts. In other words, queer identities can be seen as perpetually in a state of *becoming*, such that, rigid, persistent categorizations into states of being can actually be antithetical to these identities. Pushing towards actionable interventions, Tomasev et al. [111] suggest moving past attempts to more accurately label queer individuals and groups as a way of achieving fairness and looking instead to qualitatively engage with queer experiences with platforms and services to see how cisheteronormativity crops up in system design.

Somewhat similarly, in studies of race and racism it has been argued that race should be seen as a "a dynamic and interactive process, rather than a fixed thing that someone has" [87]. Especially for multiracial individuals, there is immense malleability in how they are perceived by others, how they perceive themselves, and what they choose to accentuate in their presentation and interactions to influence various forms of racial classification [87]. Similar to the case with queer identities, attempts to develop and enforce fairness constraints around more static, decontextualized notions of race will miss the ways in which forcing groups into static boxes is itself a form of unfairness. As such, when it is not possible to work with these fluid identity groups directly to understand how systems fail to accommodate them, data subjects should at the very least be given opportunities to update or clarify their demographics in cases where data is collected over an extended period of time and it is used in variable contexts [100].

Even in cases where groups feel adequately represented by a categorization schema, however, the categories can become harmful depending on how they are used. When demographic categories start to form the basis for differences in servicing, such as in advertising and content recommendation, there is a risk of reinforcing and naturalizing the distinctions between groups. Especially in

cases where demographic variables are uncritically adopted as an axis for differential analysis, varying outcomes across groups can be incorrectly attributed to these variables, as has occurred many times in medical research [18], which in turn reinforces the notion that the differences between groups are natural and not a result of other social factors [48]. With regards to race, a categorization schema that is conclusively not genetic or otherwise biological [76], this has been described as the risk of studying race instead of racism. By looking for differences between what groups do instead of how groups are treated, it encourages attributing responsibility to oppressed groups for their own oppression. For example, in the creation of recidivism risk scores tools for the criminal justice system, there has been extensive focus on what factors increase the accuracy of criminality prediction [7]. However, given how criminality is usually defined — namely, that an individual has been arrested and charged for a crime — the factors that end up predicting criminality most accurately are often just the factors that increase one's likelihood to be targeted by discriminatory policing [7].

6.3 Private Control Over Scoping Bias and Discrimination

As a final risk to consider, the assessment of inequality and discrimination is a not rigidly defined or widely agreed upon process. Rather, institutions that collect demographic data have a wide range of techniques and approaches they can possibly employ when it comes to both collecting data and interpreting that data. As such, if an institution is asking already marginalized groups to share information for the purposes of assessing unfairness, it is imperative for that institution to operationalize fairness in a way that is aligned with these groups' interests.

In determining what standards of fairness an institution is likely to use, it can be instructive to consider the institution's motivations for conducting measurements of fairness in the first place. Though there are many reasons an institution might try to assess and mitigate discrimination and inequalities in their machine learning and algorithmic decision-making systems, much of this work is motivated at least in part by concerns around liability [3, 53, 94]. Generally speaking, however, legal notions of discrimination and fairness remain somewhat limited, often esteeming "neutral" decision-making that attempts to treat everyone the same way as the path towards equality [118, 123, 124]. As such, most deployed methods in the algorithmic fairness space are geared towards "de-biasing" decision-making to make it more neutral, rather than trying to directly achieve equality, equity, or another form of social justice [5]. Given disparate starting points for disenfranchised groups, however, this view that neutrality can lead to a more equal world is both risky and unrealistic, as attempts to be neutral or objective often have the effect of reinforcing the status quo [38, 42]. Despite this, commitments to neutrality remain the norm for many governmental and corporate policies.

Another element of most technical approaches to fairness measurement is that they are strictly formalized. Formalizability refers to the degree to which it is possible to represent a definition of fairness through mathematical or statistical terms — for instance, defining fairness as correctly positively categorizing individuals from different groups at the same rate (i.e. equality of true positive rates [74]) is distinctly formalizable. Formalizability is an important attribute of fairness when it has to also coincide with the system design values of efficiency and scalability, because formalization enables a system designer to treat many different problems (e.g. racism, sexism, ableism) similarly. That being said, it also relies on treating much of the world as static. As Green and Viljoen [42] have argued, by treating the point of decision-making as the only possible site of intervention (i.e. adjusting predictions to adhere to some notion of fairness), these attempts at formalization hold fixed many of the engines of discrimination, such as the ways in which different groups interact with institutions and why differences might exist between groups in the first place.

Just as defining fairness, discrimination, or bias is impacted by an institution's goals and values, the collecting, processing, and interpreting of data is never truly objective. In other words, data is never "raw" because it is shaped by the conditions in which it was collected, the methods that were used, and the goals of measuring the world in the first place [40]. This brings up a salient source of risk in the collection of demographic data: the types of discrimination and inequality that can be assessed using demographic data are largely determined by what other types of data are being collected. For instance, it might be possible to detect that a risk score recidivism tool has unequal outcomes for members of different groups, but without accurate data about interactions between suspects or defendants and police and judges, it may not be possible to accurately assess why these inequalities show up in the data and thus how to best address them [9]. Given that data collection efforts must be consciously designed, data always reflects some viewpoint on what is important to understand about the world. When those collecting data have blindspots about what impacts decision-making and individuals' life experiences, various forms of discrimination and inequality run the risk of being misread as inherent qualities of groups or cultural differences between them [28]. This is why historian Khalil Gibran has argued that the seemingly objective focus on data and statistical reasoning has replaced more explicitly racist understandings of racial difference [77], a shift made possible by disaggregated data.

Taking these subjectivities of fairness measurement into account, there is a significant risk that the collection of demographic data enables private entities to selectively tweak their systems and present them as fair without meaningfully improving the experience of marginalized groups. So long as the data used to assess fairness is collected and housed by private actors, these actors are given substantial agency in scoping what constitutes fair decision-making going forward. One striking example of this already occurring is the creation and normalization of "actuarial fairness," or the notion that "each person should pay for his own risk," in the insurance industry [81]. Using statistical arguments about the uneven distribution of risk across different demographic categories, industry professionals were able to make the case for what previously might have been considered outright discrimination — charging someone more for insurance because their immutable demographic categorizations increase their statistical risk [81].

Finally, though this work is motivated by the documented unfairness of ADMS, it is critical to recognize that bias and discrimination are not the only possible harms stemming directly from ADMS. As

recent papers and reports have forcefully argued, focusing on debiasing datasets and algorithms can draw attention away from other, possibly more salient harms [5]. For many ADMS that are clearly susceptible to bias, the greater source of harm could arguably be the deployment of the system in the first place [8, 73]. Attempting to collect demographic data in these cases will likely do more harm than good, as demographic data will draw attention away from harms inherent to the system and towards seemingly resolvable issues around bias.

7 PATHS FORWARD

7.1 Anonymity, Cryptographic Privacy, and Third Party Data Management

In response to many of the measurement concerns in Section 4, there have been a range of proposals on how to feasibly collect, manage, and employ demographic data without corporations ever directly learning users' sensitive attributes. Using technical methods that enable trusted third parties to be the data collectors and holders, researchers and practitioners have found ways to maintain individual non-identifiability throughout the data's use [62, 116].

Most of these approaches prioritize anonymizing datasets that include demographics by enforcing k-anonymity, p-sensitivity, and/or differential privacy. Ensuring k-anonymity involves narrowing down the fields in a dataset and lumping variable ranges together in order to ensure that no individual has a unique set of values in the dataset that they might be re-identified with [108]. P-sensitivity involves perturbing sensitive attribute responses such that even if you know all the other variables for an individual member of the dataset, you would not be able to concretely determine the individual's sensitive attribute [13]. Differential privacy, unlike the previous two strategies, focuses instead on the model or analysis resulting from the use of sensitive data and ensures that the model or analysis would be unchanged by the removal of an individual, such that it is impossible to tell post-processing whether a specific individual was included or not in the dataset [56]. It is important to note that all of these anonymization approaches, however, add their own sources of bias that can generate misleading conclusions [46, 64].

Another range of strategies employ secure, multiparty computation (SMPC) as a means of protecting sensitive attribute data from the primary institution and outside attackers. This approach ensures individual sensitive attribute data remains encrypted at each stage of use, while still being able to carry out basic computations with the encrypted data to generate aggregate level takeaways [2, 62].

What these methods often neglect, however is the range of individual harms that extend from the relational aspects of data and the community level risks of demographic data collection. Secure computation techniques are generally concerned with an "identifiability" notion of privacy which centers the question of if data is attributable to an individual. What they miss is a more "control" centric understanding of privacy, which speaks to the ability of individuals to influence what data about them exists and how it gets used [117]. Of the risks discussed in sections 5 and 6, anonymization and cryptographic privacy can protect against the direct discrimination of having one's sensitive attributes revealed and some types of data misuse and surveillance, but they do not address concerns of (mis)representation and loss of agency. As such, we turn to recent proposals around forms of participatory data governance as a possible mechanism for mitigating these risks.

7.2 Participatory Governance of Sensitive Data

Data governance is an increasingly popular topic of discussion in light of the ever-growing swaths of data held by governments and corporations used without citizen or consumer accountability. Borrowing from Micheli et al. [71], data governance can be described as "*the power relations between all the actors affected by, or having an effect on, the way data is accessed, controlled, shared, and used, the various socio-technical arrangements set in place to generate value from data, and how such value is redistributed between actors.*" We argue that emerging data governance models which move away from individualistic data rights towards collective forms of data governance [117] can help mitigate the risks to individuals and communities described above and enable the responsible collection and use of demographic data.

While a range of alternative data governance models have been proposed [114], data cooperatives and data trusts hold notable promise for overcoming both the individual and community risks outlined in this paper. Data cooperatives are characterized by a "de-centralized data governance approach in which data subjects voluntarily pool their data together to create a common pool for mutual benefits" [51]. Data trusts, on the other hand, generally make use of a more centralized structure that relies on an independent data "trustee" to steward over the voluntary pooling and external sharing or contracting of data [1]. Notably, these forms of data governance addresses the power imbalances characteristic of current data regimes by prioritizing the distribution of access and rights to data across its members.

Looking back to the range of risks described in sections 5 and 6, these types of participatory data governance strategies can help address each in unique ways. In terms of the individual risk of privacy, participatory governance strategies inherently offer greater control to individuals over what is known about them by whom. And though nothing in the definition of data trusts or data cooperatives requires data to be nonidentifiable, both governance strategies could be used to establish the type of trusted third party relationships described in the previous section to enable encrypted or anonymized data use. In this way, we do not believe that participatory governance strategies are an alternative to cryptographic privacy and third-party data arrangements, but that they can actually be complementary strategies in mitigating privacy risks. These types of data governance models also have the potential to mitigate the risks of individual miscategorization and identity misrepresentation due to involvement of the data subjects in decisions regarding what types of data are collected and how these are categorized, and ability to raise concerns through deliberations or with data stewards. Therefore, it's possible to imagine that a participatory data governance model would include categories that more accurately represent the individuals it seeks to benefit. Similarly, the risks of data misuse or use beyond informed consent are also greatly mitigated given that it is one of the key focuses of participatory

data governance to have clear, concrete boundaries around what data is collected, why it's collected, and how it's used.

Turning to the community level risks of demographic data collection, the situation becomes more nuanced. The threat of expanding surveillance infrastructure, while reduced, is not altogether mitigated by the types of governance structures we have described or by voluntary participation in data collection efforts more broadly. Depending on how individuals are incentivized to contribute their data to a data trust or data cooperative, there is some threat of already exploited and disenfranchised populations disproportionately offering up their data for financial or material benefits, reinforcing existing disparities in privacy [68, 115]. Furthermore, surveillance operates on a community scale, not just at the level of the individual. Given the relational nature of data, data freely shared with a data trust or cooperative that is then used by private entities to build ADMS or other types of machine learning models has implications for those that have not shared their data as well. As seen through Immigration and Customs Enforcement's (ICE) purchase and use of mobile phone location data to model the movement patterns of "undocumented immigrants" [109], access to the sensitive attributes of some individuals (e.g. documentation status), can enable the surveillance of a much larger group. As such, it is incumbent upon the data trustees and data cooperatives to stay aware of how the data they wield carries risks for individuals outside their organization and to mitigate these risks where they can.

Participatory governance models can also provide a point of intervention to addressing the risk of group misrepresentation and the reinforcement of oppressive categories by enabling previously disenfranchised groups to directly define their group identity and exercise control over the applications of their community's data. One example where the utility of this has already been seen is in the field of Indigenous Data Sovereignty, which centers Indigenous peoples' prerogative to govern the collection, access, and use of their data. In cases where Indigenous tribes have lead their own data collection efforts, they have pushed back against external categorization schemas of determining tribal citizenship, such as externally imposed standards of "blood quantum," in order to more accurately define tribal membership and tell the story of their tribes [93]. Similarly, participatory governance models give data subjects the ability to exercise influence over how fairness is defined, what types of fairness assessments should be conducted, and what data should be used to conduct those assessments in a given context. This reduces the control private actors have over the operationalization of fairness and increases the alignment of fairness objectives with data subjects interests and their desired interactions with specific ADMS. However, realizing this potential will likely require some expertise around algorithmic fairness or discrimination from within the data governance structure.

While participatory data governance strategies present an opportunity to mitigate many of the risks and harms discussed in this paper, the feasibility concern from Section 4 still looms large. If the builders of ADMS are not externally compelled to approach demographic data collection and use responsibly, or even to engage with anti-discrimination in the first place, it is unlikely that data cooperatives or trusts form in order to responsibly manage and provide this data. As such, getting there will require the efforts of practitioners on the inside making the case for working with external data governance structures and of academics, policy-makers, and activists on the outside pushing for the enforcement of anti-discrimination standards and data privacy protections that would encourage corporations to only access sensitive demographic data through these limited means. As companies like Meta and AirBnB start to explore third party data holder arrangements for fairness assessments, however, this future starts to look more feasible [2, 13].

ACKNOWLEDGMENTS

We are grateful to the diverse set of individuals who engaged with us over the last year through one-on-one calls as well as the PAI-hosted FAccT CRAFT workshop and RightsCon session. We would like to thank our colleagues Christine Custis and Tina Park who provided advice and feedback on drafts of this paper. While this document reflects the input of individuals representing many PAI Partner organizations, it should not be read as representing the views of any particular organization or individual or any specific PAI Partner.

FUNDING DISCLOSURE

Funding for this study was provided by Partnership on AI. Partnership on AI is funded by a combination of philanthropic institutions and corporate charitable contributions. Primary corporate funding is always considered general operating support and legally classified as non-earmarked charitable contributions (not donations in exchange for goods or services, or quid pro quo contributions) to avoid the possibility of conflict in corporate funders having undue influence on Partnership on AI's agenda or on particular programs. More detail on Partnership on AI's funding and governance is available online.

REFERENCES

[1] Aapti Institute and Open Data Institute. 2021. *Enabling data sharing for social benefit through data trusts.* Technical Report. Global Partnership on Artificial Intelligence. https://gpai.ai/projects/data-governance/data-trusts/enabling-data-sharing-for-social-benefit-data-trusts-interim-report.pdf

[2] Rachad Alao, Miranda Bogen, Jingang Miao, Ilya Mironov, and Jonathan Tannen. 2021. *How Meta is working to assess fairness in relation to race in the U.S. across its products and systems.* Technical Report. Meta AI. 34 pages.

[3] McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. 2021. "What We Can't Measure, We Can't Understand": Challenges to Demographic Data Procurement in the Pursuit of Fairness. *arXiv:2011.02282 [cs]* (Jan. 2021). http://arxiv.org/abs/2011.02282 arXiv: 2011.02282.

[4] Chloé Bakalar, Renata Barreto, Miranda Bogen, Sam Corbett-Davies, Melissa Hall, Isabel Kloumann, Michelle Lam, Joaquin Quiñonero Candela, Manish Raghavan, Joshua Simons, Jonathan Tannen, Edmund Tong, Kate Vredenburgh, and Jiejing Zhao. 2021. Fairness On The Ground: Applying Algorithmic Fairness Approaches To Production Systems. (2021), 12.

[5] Agathe Balayn and Seda Gürses. 2021. *Beyond Debiasing.* Technical Report. European Digital Rights. https://edri.org/wp-content/uploads/2021/09/EDRi_Beyond-Debiasing-Report_Online.pdf

[6] Erin Banco and Darius Tahir. 2021. CDC under scrutiny after struggling to report Covid race, ethnicity data. https://www.politico.com/news/2021/03/09/hhs-cdc-covid-race-data-474554

[7] Chelsea Barabas. 2019. Beyond Bias: Re-Imagining the Terms of 'Ethical AI' in Criminal Law. *SSRN Electronic Journal* (2019). https://doi.org/10.2139/ssrn.3377921

[8] Chelsea Barabas. 2020. To Build a Better Future, Designers Need to Start Saying 'No'. https://onezero.medium.com/refusal-a-beginning-that-starts-with-an-end-2b055bfc14be

[9] Chelsea Barabas, JB Rubinovitz, Colin Doyle, and Karthik Dinakar. 2020. Studying Up: Reorienting the study of algorithmic fairness around issues of power. (2020), 10.

[10] Solon Barocas, Anhong Guo, Ece Kamar, Jacquelyn Krones, Meredith Ringel Morris, Jennifer Wortman Vaughan, Duncan Wadsworth, and Hanna Wallach. 2021. Designing Disaggregated Evaluations of AI Systems: Choices, Considerations, and Tradeoffs. *arXiv:2103.06076 [cs]* (March 2021). http://arxiv.org/abs/2103.06076 arXiv: 2103.06076.

[11] Solon Barocas and Karen Levy. 2019. *Privacy Dependencies.* SSRN Scholarly Paper ID 3447384. Social Science Research Network, Rochester, NY. https://papers.ssrn.com/abstract=3447384

[12] Solon Barocas and Andrew D. Selbst. 2016. Big Data's Disparate Impact. *California Law Review* 104 (2016), 671. https://heinonline.org/HOL/Page?handle=hein.journals/calr104&id=695&div=&collection=

[13] Sid Basu, Ruthie Berman, Adam Bloomston, John Cambell, Anne Diaz, Nanako Era, Benjamin Evans, Sukhada Palkar, and Skyler Wharton. 2020. *Measuring discrepancies in Airbnb guest acceptance rates using anonymized demographic data.* Technical Report. AirBnB. https://news.airbnb.com/wp-content/uploads/sites/4/2020/06/Project-Lighthouse-Airbnb-2020-06-12.pdf

[14] Ruha Benjamin. 2019. *Race After Technology: Abolitionist Tools for the New Jim Code.* Polity, Medford, MA.

[15] Rena Bivens. 2017. The gender binary will not be deprogrammed: Ten years of coding gender on Facebook. *New Media & Society* 19, 6 (June 2017), 880–898. https://doi.org/10.1177/1461444815621527 Publisher: SAGE Publications.

[16] Miranda Bogen, Aaron Rieke, and Shazeda Ahmed. 2020. Awareness in practice: tensions in access to sensitive attribute data for antidiscrimination. In *Proceedings of the 2020 conference on fairness, accountability, and transparency.* 492–500.

[17] Geoffrey C. Bowker and Susan Leigh Star. 1999. *Sorting things out: classification and its consequences.* MIT Press, Cambridge, Mass.

[18] Lundy Braun, Anne Fausto-Sterling, Duana Fullwiley, Evelynn M. Hammonds, Alondra Nelson, William Quivers, Susan M. Reverby, and Alexandra E. Shields. 2007. Racial Categories in Medical Practice: How Useful Are They? *PLOS Medicine* 4, 9 (Sept. 2007), e271. https://doi.org/10.1371/journal.pmed.0040271 Publisher: Public Library of Science.

[19] Simone Browne. 2015. *Dark Matters: On the Surveillance of Blackness.* Duke University Press. https://doi.org/10.1515/9780822375302 Publication Title: Dark Matters.

[20] Consumer Financial Protection Bureau. 2014. Using publicly available information to proxy for unidentified race and ethnicity. (2014). https://www.consumerfinance.gov/data-research/research-reports/using-publicly-available-information-to-proxy-for-unidentified-race-and-ethnicity/

[21] Moritz Büchi, Eduard Fosch-Villaronga, Christoph Lutz, Aurelia Tamò-Larrieux, and Shruthi Velidi. 2021. Making sense of algorithmic profiling: user perceptions on Facebook. *Information, Communication & Society* 0, 0 (Oct. 2021), 1–17. https://doi.org/10.1080/1369118X.2021.1989011 Publisher: Routledge _eprint: https://doi.org/10.1080/1369118X.2021.1989011.

[22] José González Cabañas, Ángel Cuevas, Aritz Arrate, and Rubén Cuevas. 2021. Does Facebook use sensitive data for advertising purposes? *Commun. ACM* 64, 1 (Jan. 2021), 62–69. https://doi.org/10.1145/3426361

[23] Hongyan Chang and Reza Shokri. 2021. On the Privacy Risks of Algorithmic Fairness. In *2021 IEEE European Symposium on Security and Privacy (EuroS P).* 292–303. https://doi.org/10.1109/EuroSP51992.2021.00028

[24] Jiahao Chen. 2018. Fair lending needs explainable models for responsible recommendation. *arXiv:1809.04684 [cs, stat]* (Sept. 2018). http://arxiv.org/abs/1809.04684 arXiv: 1809.04684.

[25] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. 2019. Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved. *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19* (2019), 339–348. https://doi.org/10.1145/3287560.3287594 arXiv: 1811.11154.

[26] Le Chen, Ruijun Ma, Anikó Hannák, and Christo Wilson. 2018. Investigating the Impact of Gender on Rank in Resume Search Engines. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18.* ACM Press, Montreal QC, Canada, 1–14. https://doi.org/10.1145/3173574.3174225

[27] Danielle Keats Citron and Daniel J. Solove. 2021. *Privacy Harms.* SSRN Scholarly Paper ID 3782222. Social Science Research Network, Rochester, NY. https://doi.org/10.2139/ssrn.3782222

[28] Roderic Crooks and Morgan Currie. 2021. Numbers will not save us: Agonistic data practices. *The Information Society* 0, 0 (May 2021), 1–19. https://doi.org/10.1080/01972243.2021.1920081 Publisher: Routledge _eprint: https://doi.org/10.1080/01972243.2021.1920081.

[29] Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2015. Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination. *Proceedings on Privacy Enhancing Technologies* 2015, 1 (April 2015), 92–112. https://doi.org/10.1515/popets-2015-0007

[30] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial Bias in Hate Speech and Abusive Language Detection Datasets. *arXiv:1905.12516 [cs]* (May 2019). http://arxiv.org/abs/1905.12516 arXiv: 1905.12516.

[31] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. *Proceedings of the International AAAI Conference on Web and Social Media* 11, 1 (May 2017), 512–515. https://ojs.aaai.org/index.php/ICWSM/article/view/14955 Number: 1.

[32] Robin Dembroff. 2018. Real Talk on the Metaphysics of Gender. *Philosophical Topics* 46, 2 (2018), 21–50. https://doi.org/10.5840/philtopics201846212

[33] Nora A Draper and Joseph Turow. 2019. The corporate cultivation of digital resignation. *New Media & Society* 21, 8 (Aug. 2019), 1824–1839. https://doi.org/10.1177/1461444819833331 Publisher: SAGE Publications.

[34] Marc N Elliott, Peter A Morrison, Allen Fremont, Daniel F McCaffrey, Philip Pantoja, and Nicole Lurie. 2009. Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Services and Outcomes Research Methodology* 9, 2 (2009), 69. Publisher: Springer.

[35] Virginia Eubanks. 2017. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor.* St. Martin's Press, New York, NY.

[36] European Parliament and Council of European Union. 2016. Regulation (EU) 2016/679 (General Data Protection Regulation). https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN

[37] Tom Farrand, Fatemehsadat Mireshghallah, Sahib Singh, and Andrew Trask. 2020. Neither Private Nor Fair: Impact of Data Imbalance on Utility and Fairness in Differential Privacy. In *Proceedings of the 2020 Workshop on Privacy-Preserving Machine Learning in Practice (PPMLP'20).* Association for Computing Machinery, New York, NY, USA, 15–19. https://doi.org/10.1145/3411501.3419419

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

[39] E. Fosch-Villaronga, A. Poulsen, R. A. Søraa, and B. H. M. Custers. 2021. A little bird told me your gender: Gender inferences in social media. *Information Processing & Management* 58, 3 (May 2021), 102541. https://doi.org/10.1016/j.ipm.2021.102541

[40] Lisa Gitelman. 2013. *Raw Data Is an Oxymoron.* MIT Press. Google-Books-ID: Be5ZAgAAQBAJ.

[41] Bryce W Goodman. 2016. A step towards accountable algorithms? algorithmic discrimination and the european union general data protection. In *29th conference on neural information processing systems (NIPS 2016), barcelona. NIPS foundation.*

[42] Ben Green and Salomé Viljoen. 2020. Algorithmic realism: expanding the boundaries of algorithmic thought. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.* ACM, Barcelona Spain, 19–31. https://doi.org/10.1145/3351095.3372840

[43] David Grossman. 2018. Amazon Fired Its Resume-Reading AI for Sexism. https://www.popularmechanics.com/technology/robots/a23708450/amazon-resume-ai-sexism/ Section: Robots.

[44] Kevin Guyan. 2022. *Queer Data: Using Gender, Sex and Sexuality Data for Action.* Bloomsbury Publishing.

[45] Ian Hacking. 1995. The looping effects of human kinds. In *Causal cognition: A multidisciplinary debate.* Clarendon Press/Oxford University Press, New York, NY, US, 351–394.

[46] Sara Hajian and Josep Domingo-Ferrer. 2012. A Study on the Impact of Data Anonymization on Anti-discrimination. In *2012 IEEE 12th International Conference on Data Mining Workshops.* IEEE, Brussels, Belgium, 352–359. https://doi.org/10.1109/ICDMW.2012.19

[47] Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M. Branham. 2018. Gender Recognition or Gender Reductionism?: The Social Implications of Embedded Gender Recognition Systems. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18.* ACM Press, Montreal QC, Canada, 1–13. https://doi.org/10.1145/3173574.3173582

[48] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a Critical Race Methodology in Algorithmic Fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.* ACM, Barcelona Spain, 501–512. https://doi.org/10.1145/3351095.3372826

[49] Alex Hern. 2018. Google's solution to accidental algorithmic racism: ban gorillas. *The Guardian* (Jan. 2018). https://www.theguardian.com/technology/2018/jan/12/google-racism-ban-gorilla-black-people

[50] Kashmir Hill. 2020. The Secretive Company That Might End Privacy as We Know It. https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html

[51] Chih-Hsing Ho and Tyng-Ruey Chuang. 2019. Governance of Communal Data Sharing. *Good Data* (2019), 202–219.

[52] Anna Lauren Hoffmann. 2020. Terms of inclusion: Data, discourse, violence. *New Media & Society* (Sept. 2020), 146144482095872. https://doi.org/10.1177/1461444820958725

[53] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudík, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need? *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19* (2019), 1–16. https://doi.org/10.1145/

3290605.3300830 arXiv: 1812.05239.

[54] Lily Hu. 2020. What is 'Race' in Algorithmic Discrimination on the Basis of Race? https://www.youtube.com/watch?v=m5GdAgZnpzA

[55] Thomas Hupperich, Dennis Tatang, Nicolai Wilkop, and Thorsten Holz. 2018. An Empirical Study on Online Price Differentiation. In *Proceedings of the Eighth ACM Conference on Data and Application Security and Privacy (CODASPY '18)*. Association for Computing Machinery, New York, NY, USA, 76–83. https://doi.org/10.1145/3176258.3176338

[56] Matthew Jagielski, Michael Kearns, Jieming Mao, Alina Oprea, Aaron Roth, Saeed Sharifi Malvajerdi, and Jonathan Ullman. 2019. Differentially Private Fair Learning. In *Proceedings of the 36th International Conference on Machine Learning*. PMLR, 3000–3008. https://proceedings.mlr.press/v97/jagielski19a.html ISSN: 2640-3498.

[57] Jon Keegan. 2021. Facebook Got Rid of Racial Ad Categories. Or Did It? https://themarkup.org/citizen-browser/2021/07/09/facebook-got-rid-of-racial-ad-categories-or-did-it Section: Citizen Browser.

[58] Stephanie Kelley, Anton Ovchinnikov, David R. Hardoon, and Adrienne Heinrich. 2021. *Anti-discrimination Laws, AI, and Gender Bias: A Case Study in Non-mortgage Fintech Lending*. SSRN Scholarly Paper ID 3719577. Social Science Research Network, Rochester, NY. https://doi.org/10.2139/ssrn.3719577

[59] Os Keyes. 2018. The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (Nov. 2018), 88:1–88:22. https://doi.org/10.1145/3274357

[60] Os Keyes. 2019. Counting the Countless. https://reallifemag.com/counting-the-countless/

[61] Os Keyes, Zoë Hitzig, and Mwenza Blell. 2021. Truth from the machine: artificial intelligence and the materialization of identity. *Interdisciplinary Science Reviews* 46, 1-2 (April 2021), 158–175. https://doi.org/10.1080/03080188.2020.1840224 Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/03080188.2020.1840224.

[62] Niki Kilbertus, Adrià Gascón, Matt J. Kusner, Michael Veale, Krishna P. Gummadi, and Adrian Weller. 2018. Blind Justice: Fairness with Encrypted Sensitive Attributes. *arXiv:1806.03281 [cs, stat]* (June 2018). http://arxiv.org/abs/1806.03281 arXiv: 1806.03281.

[63] Jennifer King. 2019. " Becoming Part of Something Bigger" Direct to Consumer Genetic Testing, Privacy, and Personal Disclosure. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–33.

[64] Satya Kuppam, Ryan Mckenna, David Pujol, Michael Hay, Ashwin Machanavajjhala, and Gerome Miklau. 2020. Fair Decision Making using Privacy-Protected Data. *arXiv:1905.12744 [cs]* (Jan. 2020). http://arxiv.org/abs/1905.12744 arXiv: 1905.12744.

[65] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness. In *Advances in Neural Information Processing Systems*, Vol. 30. Curran Associates, Inc. https://papers.nips.cc/paper/2017/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html

[66] Kalev Leetaru. 2018. Facebook As The Ultimate Government Surveillance Tool? https://www.forbes.com/sites/kalevleetaru/2018/07/20/facebook-as-the-ultimate-government-surveillance-tool/ Section: AI & Big Data.

[67] Nancy López and Howard Hogan. 2021. What's Your Street Race? The Urgency of Critical Race Theory and Intersectionality as Lenses for Revising the U.S. Office of Management and Budget Guidelines, Census and Administrative Data in Latinx Communities and Beyond. *Genealogy* 5, 3 (Sept. 2021), 75. https://doi.org/10.3390/genealogy5030075 Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

[68] Alice E. Marwick and Danah Boyd. 2018. Privacy at the Margins| Understanding Privacy at the Margins—Introduction. *International Journal of Communication* 12, 0 (March 2018), 9. https://ijoc.org/index.php/ijoc/article/view/7053 Number: 0.

[69] Douglas S. Massey, Jacob S. Rugh, Justin P. Steil, and Len Albright. 2016. Riding the Stagecoach to Hell: A Qualitative Analysis of Racial Discrimination in Mortgage Lending. *City & Community* 15, 2 (2016), 118–136. https://doi.org/10.1111/cico.12179 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cico.12179.

[70] Paola Mavriki and Maria Karyda. 2019. Automated data-driven profiling: threats for group privacy. *Information & Computer Security* 28, 2 (Jan. 2019), 183–197. https://doi.org/10.1108/ICS-04-2019-0048 Publisher: Emerald Publishing Limited.

[71] Marina Micheli, Marisa Ponti, Max Craglia, and Anna Berti Suman. 2020. Emerging models of data governance in the age of datafication. *Big Data & Society* 7, 2 (July 2020), 2053951720948087. https://doi.org/10.1177/2053951720948087 Publisher: SAGE Publications Ltd.

[72] Jakub Mikians, László Gyarmati, Vijay Erramilli, and Nikolaos Laoutaris. 2013. Crowd-assisted search for price discrimination in e-commerce: first results. In *Proceedings of the ninth ACM conference on Emerging networking experiments and technologies (CoNEXT '13)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/2535372.2535415

[73] Yeshimabeit Milner. 2019. Abolish Big Data. https://medium.com/@YESHICAN/abolish-big-data-ad0871579a41

[74] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2020. Algorithmic Fairness: Choices, Assumptions, and Definitions. https://doi.org/10.1146/annurev-statistics-042720-125902 Archive Location: world Publisher: Annual Reviews.

[75] Brent Mittelstadt. 2017. From Individual to Group Privacy in Big Data Analytics. *Philosophy & Technology* 30, 4 (Dec. 2017), 475–494. https://doi.org/10.1007/s13347-017-0253-7

[76] Ann Morning. 2014. Does Genomics Challenge the Social Construction of Race?:. *Sociological Theory* (Oct. 2014). https://doi.org/10.1177/0735275114550881 Publisher: SAGE PublicationsSage CA: Los Angeles, CA.

[77] Khalil Gibran Muhammad. 2019. *The Condemnation of Blackness: Race, Crime, and the Making of Modern Urban America, With a New Preface*. Harvard University Press. Google-Books-ID: gqacDwAAQBAJ.

[78] Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, Ioannis Kompatsiaris, Katharina Kinder-Kurlanda, Claudia Wagner, Fariba Karimi, Miriam Fernandez, Harith Alani, Bettina Berendt, Tina Kruegel, Christian Heinze, Klaus Broelemann, Gjergji Kasneci, Thanassis Tiropanis, and Steffen Staab. 2020. Bias in data-driven artificial intelligence systems—An introductory survey. *WIREs Data Mining and Knowledge Discovery* 10, 3 (May 2020). https://doi.org/10.1002/widm.1356

[79] Jonathan A. Obar. 2020. Sunlight alone is not a disinfectant: Consent and the futility of opening Big Data black boxes (without assistance). *Big Data & Society* 7, 1 (Jan. 2020), 2053951720935615. https://doi.org/10.1177/2053951720935615 Publisher: SAGE Publications Ltd.

[80] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (Oct. 2019), 447–453. https://doi.org/10.1126/science.aax2342

[81] Rodrigo Ochigame, Chelsea Barabas, Karthik Dinakar, Madars Virza, and Joichi Ito. 2018. Beyond Legitimation: Rethinking Fairness, Interpretability, and Accuracy in Machine Learning. *International Conference on Machine Learning* (2018), 6.

[82] Anne Oeldorf-Hirsch and Jonathan A. Obar. 2019. Overwhelming, Important, Irrelevant: Terms of Service and Privacy Policy Reading among Older Adults. In *Proceedings of the 10th International Conference on Social Media and Society (SMSociety '19)*. Association for Computing Machinery, New York, NY, USA, 166–173. https://doi.org/10.1145/3328529.3328557

[83] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. *Frontiers in Big Data* 2 (July 2019), 13. https://doi.org/10.3389/fdata.2019.00013

[84] Partnership on AI. 2020. *Algorithmic Risk Assessment and COVID-19: Why PATTERN Should Not Be Used*. Technical Report. Partnership on AI. http://partnershiponai.org/wp-content/uploads/2021/07/Why-PATTERN-Should-Not-Be-Used.pdf

[85] Partnership on AI. 2022. Welcome to the Artificial Intelligence Incident Database. https://incidentdatabase.ai/

[86] Samir Passi and Solon Barocas. 2019. Problem Formulation and Fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 39–48. https://doi.org/10.1145/3287560.3287567

[87] Kristin Pauker, Chanel Meyers, Diana T. Sanchez, Sarah E. Gaither, and Danielle M. Young. 2018. A review of multiracial malleability: Identity, categorization, and shifting racial attitudes. *Social and Personality Psychology Compass* 12, 6 (2018), e12392. https://doi.org/10.1111/spc3.12392 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/spc3.12392.

[88] OiYan A. Poon, Jude Paul Matias Dizon, and Dian Squire. 2017. Count Me In!: Ethnic Data Disaggregation Advocacy, Racial Mattering, and Lessons for Racial Justice Coalitions. *JCSCORE* 3, 1 (April 2017), 91–124. https://doi.org/10.15763/issn.2642-2387.2017.3.1.91-124 Number: 1.

[89] Lincoln Quillian, John J Lee, and Mariana Oliver. 2020. Evidence from Field Experiments in Hiring Shows Substantial Additional Racial Discrimination after the Callback. *Social Forces* 99, 2 (Nov. 2020), 732–759. https://doi.org/10.1093/sf/soaa026

[90] Lincoln Quillian, Devah Pager, Ole Hexel, and Arnfinn H. Midtbøen. 2017. Meta-analysis of field experiments shows no change in racial discrimination in hiring over time. *Proceedings of the National Academy of Sciences* 114, 41 (Oct. 2017), 10870–10875. https://doi.org/10.1073/pnas.1706255114 Publisher: National Academy of Sciences Section: Social Sciences.

[91] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2019. Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices. *arXiv:1906.09208 [cs]* (Dec. 2019). https://doi.org/10.1145/3351095.3372828 arXiv: 1906.09208.

[92] Stephanie Carroll Rainie, Tahu Kukutai, Maggie Walter, Oscar Luis Figueroa-Rodríguez, Jennifer Walker, and Per Axelsson. 2019. Indigenous data sovereignty. (2019). Publisher: African Minds and the International Development Research Centre (IDRC).

[93] Stephanie Carroll Rainie, Jennifer Lee Schultz, Eileen Briggs, Patricia Riggs, and Nancy Lynn Palmanteer-Holder. 2017. Data as a Strategic Resource: Self-determination, Governance, and the Data Challenge for Indigenous Nations in the United States. *International Indigenous Policy Journal* 8, 2 (March 2017). https://doi.org/10.18584/iipj.2017.8.2.1

[94] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. 2021. Where Responsible AI meets Reality: Practitioner Perspectives on Enablers for shifting Organizational Practices. *arXiv:2006.12358 [cs]* (March 2021). https://doi.org/10.1145/3449081 arXiv: 2006.12358.

[95] Paola Ricaurte. 2019. Data Epistemologies, Coloniality of Power, and Resistance. *Television & New Media* (2019), 16.

[96] Kaili Rimfeld and Margherita Malanchini. 2020. The A-Level and GCSE scandal shows teachers should be trusted over exams results. https://inews.co.uk/opinion/a-level-gcse-results-trust-teachers-exams-592499

[97] Eva Rosen, Philip M. E. Garboden, and Jennifer E. Cossyleon. 2021. Racial Discrimination in Housing: How Landlords Use Algorithms and Home Visits to Screen Tenants. *American Sociological Review* 86, 5 (Oct. 2021), 787–822. https://doi.org/10.1177/00031224211029618 Publisher: SAGE Publications Inc.

[98] Wendy D. Roth. 2016. The multiple dimensions of race. *Ethnic and Racial Studies* 39, 8 (June 2016), 1310–1338. https://doi.org/10.1080/01419870.2016.1140793

[99] Alan Z. Rozenshtein. 2018. *Surveillance Intermediaries.* SSRN Scholarly Paper ID 2935321. Social Science Research Network, Rochester, NY. https://papers.ssrn.com/abstract=2935321

[100] Bonnie Ruberg and Spencer Ruelos. 2020. Data for queer lives: How LGBTQ gender and sexuality identities challenge norms of demographics. *Big Data & Society* 7, 1 (Jan. 2020), 2053951720933286. https://doi.org/10.1177/2053951720933286 Publisher: SAGE Publications Ltd.

[101] Aliya Saperstein. 2012. Capturing complexity in the United States: which aspects of race matter and when? *Ethnic and Racial Studies* 35, 8 (Aug. 2012), 1484–1502. https://doi.org/10.1080/01419870.2011.607504 Publisher: Routledge _eprint: https://doi.org/10.1080/01419870.2011.607504.

[102] Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R. Brubaker. 2020. How We've Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (May 2020), 1–35. https://doi.org/10.1145/3392866

[103] 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.* ACM Press, Atlanta, GA, USA, 59–68. https://doi.org/10.1145/3287560.3287598

[104] Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, Irene Y. Chen, and Marzyeh Ghassemi. 2021. CheXclusion: Fairness gaps in deep chest X-ray classifiers. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* 26 (2021), 232–243.

[105] Riti Shimkhada, A. J. Scheitler, and Ninez A. Ponce. 2021. Capturing Racial/Ethnic Diversity in Population-Based Surveys: Data Disaggregation of Health Data for Asian American, Native Hawaiian, and Pacific Islanders (AANHPIs). *Population Research and Policy Review* 40, 1 (Feb. 2021), 81–102. https://doi.org/10.1007/s11113-020-09634-3

[106] Dean Spade. 2015. *Normal Life: Administrative Violence, Critical Trans Politics, and the Limits of Law.* Duke University Press. https://doi.org/10.1515/9780822374794 Publication Title: Normal Life.

[107] Luke Stark and Jevan Hutson. 2021. *Physiognomic Artificial Intelligence.* SSRN Scholarly Paper ID 3927300. Social Science Research Network, Rochester, NY. https://doi.org/10.2139/ssrn.3927300

[108] Latanya Sweeney. 2002. k-Anonymity: A Model For Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 05 (Oct. 2002), 557–570. https://doi.org/10.1142/S0218488502001648 Publisher: World Scientific Publishing Co.

[109] Byron Tau and Michelle Hackman. 2020. Federal Agencies Use Cellphone Location Data for Immigration Enforcement. *Wall Street Journal* (Feb. 2020). https://www.wsj.com/articles/federal-agencies-use-cellphone-location-data-for-immigration-enforcement-11581078600

[110] Linnet Taylor. 2021. Public Actors Without Public Values: Legitimacy, Domination and the Regulation of the Technology Sector. *Philosophy & Technology* (Jan. 2021). https://doi.org/10.1007/s13347-020-00441-4

[111] Nenad Tomasev, Kevin R. McKee, Jackie Kay, and Shakir Mohamed. 2021. Fairness for Unobserved Characteristics: Insights from Technological Impacts on Queer Communities. *arXiv:2102.04257 [cs]* (Feb. 2021). http://arxiv.org/abs/2102.04257 arXiv: 2102.04257.

[112] UK Information Commissioner's Office. 2020. What do we need to do to ensure lawfulness, fairness, and transparency in AI systems? https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/guidance-on-ai-and-data-protection/what-do-we-need-to-do-to-ensure-lawfulness-fairness-and-transparency-in-ai-systems/ Publisher: ICO.

[113] U.S. Department of Justice. 2019. The First Step Act of 2018 : Risk and Needs Assessment System.

[114] Jonathan Van Geuns, Ana Brandusescu, and Mozilla Insights. 2020. *What Does it Mean? | Shifting Power Through Data Governance.* Technical Report. Mozilla Foundation. https://foundation.mozilla.org/en/data-futures-lab/data-for-empowerment/shifting-power-through-data-governance/

[115] Jessica Vasquez-Tokos and Priscilla Yamin. 2021. The racialization of privacy: racial formation as a family affair. *Theory and Society* 50, 5 (Aug. 2021), 717–740. https://doi.org/10.1007/s11186-020-09427-9

[116] Michael Veale and Reuben Binns. 2017. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society* 4, 2 (Dec. 2017), 205395171774353. https://doi.org/10.1177/2053951717743530

[117] Salome Viljoen. 2020. *Democratic Data: A Relational Theory For Data Governance.* SSRN Scholarly Paper ID 3727562. Social Science Research Network, Rochester, NY. https://doi.org/10.2139/ssrn.3727562

[118] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2021. *Bias Preservation in Machine Learning: The Legality of Fairness Metrics Under EU Non-Discrimination Law.* SSRN Scholarly Paper ID 3792772. Social Science Research Network, Rochester, NY. https://doi.org/10.2139/ssrn.3792772

[119] Maggie Walter. 2020. Delivering Indigenous Data Sovereignty. https://www.youtube.com/watch?v=NCsCZJ8ugPA

[120] Gregory M. Walton, David Paunesku, and Carol S. Dweck. 2012. Expandable selves. In *Handbook of self and identity, 2nd ed.* The Guilford Press, New York, NY, US, 141–154.

[121] Anne L. Washington. 2018. How to Argue with an Algorithm: Lessons from the COMPAS-ProPublica Debate. *Colorado Technology Law Journal* 17 (2018), 131. https://heinonline.org/HOL/Page?handle=hein.journals/jtelhtel17&id=145&div=&collection=

[122] Brooke Foucault Welles. 2014. On minorities and outliers: The case for making Big Data small. *Big Data & Society* 1, 1 (April 2014), 2053951714540613. https://doi.org/10.1177/2053951714540613 Publisher: SAGE Publications Ltd.

[123] Raphaële Xenidis. 2021. Tuning EU Equality Law to Algorithmic Discrimination: Three Pathways to Resilience. *Maastricht Journal of European and Comparative Law* 27 (Jan. 2021), 1023263X2098217. https://doi.org/10.1177/1023263X20982173

[124] Alice Xiang. 2021. Reconciling legal and technical approaches to algorithmic bias. *Tennessee Law Review* 88, 3 (2021).

[125] Lu Zhang, Yongkai Wu, and Xintao Wu. 2017. A Causal Framework for Discovering and Removing Direct and Indirect Discrimination. *Proceedings of the 26th International Joint Conference on Artificial Intelligence* (2017), 3929–3935. https://www.ijcai.org/proceedings/2017/549

[126] Tukufu Zuberi and Eduardo Bonilla-Silva. 2008. *White Logic, White Methods: Racism and Methodology.* Rowman & Littlefield Publishers. Google-Books-ID: WTBtAAAAQBAJ.