WHITE PAPER

Fairer Algorithmic Decision Making & Its Consequences

Interrogating the Risks and Benefits of Demographic Data Collection, Use, and Non-Use

DECEMBER 2, 2021

PARTNERSHIP ON AI

Contents

Introduction 3

Background 4

Social Risks of Non-Use 6

Hidden Discrimination 6

“Colorblind” Decision-making 7

Invisibility to Institutions of Importance 8

Social Risks of Use 10

Risks to Individuals 10

Encroachments on privacy and personal life 10

Individual Misrepresentation 12

Data misuse and use beyond informed consent 13

Risks to Communities 15

Expanding Surveillance Infrastructure in the Pursuit of Fairness 15

Misrepresentation and Reinforcing Oppressive or Overly Prescriptive Categories 15

Private Control over Scoping Bias and Discrimination 18

Conclusion 21

Acknowledgements 22

Notes 23

Introduction

Algorithmic decision-making has been widely accepted as a novel approach to overcoming the purported cognitive and subjective limitations of human decision makers by providing "objective" data-driven recommendations. Yet, as organizations adopt algorithmic decision-making systems (ADMS), countless examples of algorithmic discrimination continue to emerge. Harmful biases have been found in algorithmic decision-making systems in contexts such as healthcare, hiring, criminal justice, and education, prompting increasing social concern regarding the impact these systems are having on the wellbeing and livelihood of individuals and groups across society. In response, algorithmic fairness strategies attempt to understand how ADMS treat certain individuals and groups, often with the explicit purpose of detecting and mitigating harmful biases.

Many current algorithmic fairness techniques require access to data on a "sensitive attribute" or "protected category" (such as race, gender, or sexuality) in order to make performance comparisons and standardizations across groups. These demographic-based algorithmic fairness techniques assume that discrimination and social inequality can be overcome with clever algorithms and collection of the requisite data, removing broader questions of governance and politics from the equation. This paper seeks to challenge this assumption, 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. This involves acknowledging the value assumptions and trade-offs associated with the use and non-use of demographic data in algorithmic systems. To advance this discussion, this white paper provides a preliminary perspective on these trade-offs derived from workshops and conversations with experts in industry, academia, government, and advocacy organizations as well as literature across relevant domains. In doing so, we hope that readers will better understand the affordances and limitations of using demographic data to detect and mitigate discrimination in institutional decision-making more broadly.

If you have any feedback on this white paper or if you would like to receive updates about future demographic data research, please reach out to Sarah Villeneuve (sarah.v@partnershiponai.org) and McKane Andrus (mckane@partnershiponai.org).

Background

Demographic-based algorithmic fairness techniques presuppose the availability of data on sensitive attributes or protected categories. However, previous research has highlighted that data on demographic categories, such as race and sexuality, are often unavailable due to a range of organizational challenges, legal barriers, and practical concerns.[1] Some privacy laws, such as the EU's GDPR, not only require data subjects to provide meaningful consent when their data is collected, but also prohibit the collection of sensitive data such as race, religion, and sexuality. 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. 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.[2]

In cases where sensitive demographic data can be collected, organizations must navigate a number of practical challenges throughout its procurement. For many organizations, sensitive demographic data is collected through self-reporting mechanisms. However, self-reported data is often incomplete, unreliable, and unrepresentative, due in part to a lack of incentives for individuals to provide accurate and full information.[3] In some cases, practitioners choose to infer protected categories of individuals based on proxy information, a method which is largely inaccurate. Organizations also face difficulty capturing unobserved characteristics, such as disability, sexuality, and religion, as these categories are frequently missing and often unmeasurable.[4] Overall, deciding on how to classify and categorize demographic data is an ongoing challenge, as demographic categories continue to shift and change over time and between contexts. Once demographic data is collected, antidiscrimination law and policies largely inhibit organizations from using this data since knowledge of sensitive categories opens the door to legal liability if discrimination is uncovered without a plan to successfully mitigate it.[5]

Deciding on how to classify & categorize demographic data is an ongoing challenge, as demographic categories continue to shift & change.

In the face of these barriers, corporations looking to apply demographic-based algorithmic fairness techniques have called for guidance on how to responsibly collect and use demographic data. However, prescribing 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.[6] Therefore, developing guidance requires a deeper understanding of the risks and trade-offs inherent to the use and non-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.

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 is (1) often misguided because proposed debiasing methods are only relevant for a subset of the kinds of bias ADMS introduce or reinforce, and (2) likely to draw attention away from other, possibly more salient harms.[7] In the first case, harms from tools such as recommendation systems, content moderation systems, and computer vision systems might be characterized as a result of various forms of bias, but resolving bias in those systems generally involves adding in more context to better understand differences between groups, not just trying to treat groups more similarly. In the second case, there are many ADMS that are clearly susceptible to bias, yet the greater source of harm could arguably be the deployment of the system in the first place. Pre-trial detention risk scores provide one such example. Using statistical correlations to determine if someone should be held without bail, or, in other words, potentially punishing individuals for attributes outside of their control and past decisions unrelated to what they are currently being charged for, is itself a significant deviation from legal standards and norms, yet most of the debate has focused around how biased the predictions are. 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.

Social Risks of Non-Use

When demographic data is discussed in the context of algorithmic decision-making, it is most often to make the case for why disaggregated data is necessary to make fairer decisions. In this section, "Social Risks of Non-Use," we identify some key features of what the addition of demographic data does for efforts to detect and mitigate discrimination in institutional decision-making more broadly. In the next section, "Social Risks of Use," we will delve into some of the less frequently considered risks of actually collecting and using this data.

Hidden Discrimination

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 most cases, when ADMS discriminate against protected groups, they do so indirectly. 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 have the effect of discriminating across groups.

In order to understand how algorithms can discriminate, it is important to consider the different ways in which bias can enter the picture. The first point of entry is most obviously 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.[8]

If optimizing for a goal that is poorly defined, it is likely that a system will reproduce historical inequity & discrimination, just under a guise of objectivity.

Using biased data, however, is not the only way that ADMS can have a discriminatory impact. How ADMS are designed and towards what kinds of objectives have a large bearing on how discriminatory their outcomes are. 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. For example, the UK higher education admission algorithm attempted to define aptitude as a combination of a *predicted performance* and *secondary school quality*, systematically biasing the outcomes for those coming from poorer or less-established secondary schools.[9] 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.[10]

Though the types of discrimination discussed here represent a small subset of the myriad ways that ADMS can discriminate, we are still confronted with a difficult question — how should practitioners assess all 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, but almost all of the proposed methods require that the datasets in use include the potentially discriminated against demographic attributes. Generally speaking, however, prior work has shown that demographic attributes are only collected once a narrow, enforceable definition of discrimination is codified into law or corporate standards.[11] Furthermore, the issue of missing demographic data is often only confronted and explicitly addressed once assessment and/or enforcement efforts begin in earnest.[12] Even then, 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).

As such, we frequently see a cycle of ADMS development and deployment, exposure of egregious discrimination through individual reports,* and then ad hoc system redesigns. Without access to demographic attributes, it's 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.

* For example, when the Google Photos app automatically tagged images of **Black users as gorillas** or when the Apple Card reportedly offered **lower credit limits to women**. Both of these issues were uncovered by users publicly sharing their experiences on social media, a relatively common way that algorithmic mishaps get exposed and end up on PAI's **AI Incidents Database**.

"Colorblind" Decision-making

Just as an absence of demographic data can prevent practitioners from uncovering various forms of social or institutional discrimination, it can also prevent them from making systems that have the explicit goal of addressing historical discrimination. In fact, under a number of legal and policy frameworks, ignoring or omitting demographic attributes altogether is actually considered non-discriminatory. 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.[13] Often, this is because the decision-making systems we build take in historical data and learn to reproduce historical biases embedded in that data. Sometimes this happens because the system explicitly learns to prioritize accuracy or performance for one group over another by using "proxies" for demographic attributes (e.g., pregnancy status is often a proxy for gender). By cobbling together attributes such as zip code, income, parental status, etc., machine learning systems can "reconstruct" demographic category membership, if doing so is beneficial to the prediction task

at hand.[14] In other cases, discrimination stems from prioritizing certain attributes that exhibit disparities across groups as a result of historical oppression, such as wealth or educational attainment.

Addressing these forms of discrimination, however, is not so easy as just introducing demographic data to the dataset. As seen in the debates around the COMPAS recidivism prediction algorithm, fairness or discrimination can be defined in many, often conflicting, ways.[15] This raises a second type of unawareness or color-blindness that is more insidious: the belief that if a decision is not made because of a demographic attribute or some proxy thereof, that the decision cannot be discriminatory. For example, credit-scoring institutions now make use of data that is much more closely linked to race and other demographic categories than the concept of "credit-worthiness," such as criminal history and how one communicates online.[16] Looking specifically at criminal history, a social constructivist perspective on race would suggest that being subjected to discriminatory (if not outright predatory) policing is part and parcel of what it means to be categorized as Black in the United States.[17] As such, when we treat demographic categories as standalone attributes and blind ourselves to the web of relationships that constitute a demographic category, we espouse a worldview that we should not consider systemically rooted differences across groups, individualizing the responsibility for historical disenfranchisement, oppression, and inequality. As has been thoroughly explored in other work and domains, attempting to ignore societal differences across demographic groups often works to reinforce or reproduce systems of oppression.[18] Within the algorithmic decision-making space specifically, Eubanks[19] has referred to this approach as reinforcing "feedback loops of injustice," where systemic inequalities are reflected in data that is then used to make "objective" decisions that deepen the inequalities.

Attempting to ignore societal differences across demographic groups often works to reinforce or reproduce systems of oppression.

Thus, while there is potential benefit to collecting demographic data to enable more "attribute aware" decision-making, corporations and public institutions must be committed to addressing historical discrimination and oppression to realize this benefit. The Toolkit for Centering Racial Equity Throughout Data Integration covers in more detail than we can here what patterns of discrimination might be reproduced through the use of historical data and/or algorithmic decision-making.

Invisibility to Institutions of Importance

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. Conversely, when data

collection efforts omit certain demographic categories, or even demographics entirely, groups can be rendered invisible to the institutions relying on this data. 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.[20] 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.[21]

The risk of some groups being rendered invisible, however, can be further heightened as institutions turn to inferring demographic attributes instead of collecting them from data subjects directly. Common techniques used by public and private institutions, such as Bayesian Improved Surname Geocoding (BISG), which uses an individual's name and zip code to predict their race,[22] often rely on a very limited set of demographic categories that obscure subgroups that might need more specialized treatment. For example, there have been many efforts to distinguish between Asian American and Pacific Islander (AAPI) populations in health[23] and education[24] 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). Similarly, we have seen how inferring genders for the purposes of content recommendation and advertising can misinterpret or outright ignore individuals of minoritized gender identities.[25] As such, when increasing group visibility is a salient reason for collecting demographic data, it is critical that such data is collected with the involvement and consent of members of that group.

It is important to note, however, 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. Furthermore, just because a group is made visible by disaggregated data, it does not follow that the institutions making use of the data are committed to better tailoring their systems to the needs of that group. As we have seen time and time again with the hyper-surveillance of Black and Brown communities in the United States by law enforcement and public service agencies, some initial visibility can be used to justify more and more invasive forms of visibility.[26]

Social Risks of Use

When demographic data is used, it carries risks for both individuals and groups. Here, we discuss both the affordances and limitations of using demographic data to detect and mitigate discrimination in institutional decision-making more broadly. Our goal is not to suggest that demographic data shouldn't be used, but rather to build out a clearer picture of what it is we are trying to use it for so as to outline the minimum conditions we expect our demographic data governance strategies to enable.

Risks to Individuals

ENCROACHMENTS ON PRIVACY AND PERSONAL LIFE

Likely the first concern that many would have when it comes to collecting or using sensitive demographic data are the risks from breaching individual privacy. Demographic attributes such as race, ethnicity, country of birth, gender, and sexuality are rarely inconsequential aspects of one's identity that can be shared or learned without risk. Quite to the contrary, sharing or otherwise determining these 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.[27] As such, we should anticipate that any current efforts to collect sensitive demographic attributes will at some point in the pipeline require tying the attributes to individuals, risking individuals' privacy.

One clear privacy risk of obtaining an individual's demographics is that these attributes are still the basis for many types of discrimination.

One clear privacy risk of obtaining an individual's demographics is that these attributes are still 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. In domains such as hiring,[28] advertising,[29] and pricing,[30] direct forms of discrimination, algorithmically mediated or not, are relatively common. For domains like advertising, discriminatory practices are often justified by claims that differential treatment results in better 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.[31]

In the most pernicious cases, demographic attributes 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 sensitive attribute (e.g., documentation status) is not collected, other collected attributes (e.g., country of birth and spoken language) can be used to help infer the targeted attribute. As corporate data becomes increasingly requested by and made available to state agencies,[32] it is critical that practitioners consider what types of identity-based violence individuals might be exposed to by sharing certain attributes.

A commonly suggested approach to reducing these forms of direct targeting risk is to "anonymize" or "de-identify" datasets. Experimental methods, however, have achieved high "re-identification" accuracy for datasets with numerous demographic attributes.[33] Marginalized individuals are especially vulnerable to these types of re-identification strategies, as there tend to be fewer data subjects in datasets that share their demographic attributes. Attempting to address this problem, researchers have proposed various differential privacy techniques for ensuring both a technical definition of fairness and non-identifiability, but these approaches are experimental and can inhibit other types of demographic analysis.[34]

Finally, another salient privacy risk to consider 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 attributes 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."*[35] Barocas and Levy[37] 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.[38]

* When these categories are explicitly labeled, it can result in the type of backlash that Facebook faced for including inferred "ethnic affinity" in their ad-targeting categories.[36]

For many of these privacy risks, we might expect privacy regulation such as the GDPR or California's 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.[39] As we just discussed, an individual's sensitive attributes need not be explicitly collected or inferred in order for algorithmic systems to treat them as part of a specific group. 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."[40] In other words, users are resigned to provide 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.[41] 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.

INDIVIDUAL MISREPRESENTATION

In an effort to mitigate bias, some organizations seek to make their datasets more "representative" by including more data on different demographic categories such as race and gender. However, this is often done without a deeper engagement with the categories themselves or the collection methods used. How demographic data is coded and represented in datasets — specifically, what categories are being used to define individual characteristics — can have an enormous impact on the representation of marginalized individuals. In the context of ADMS, individual misrepresentation can lead to discrimination and disparate impacts.

How demographic data is coded & represented in datasets can have an enormous impact on the representation of marginalized individuals.

Gender and race are two demographic categories that have long and complex socio-political histories of classification. Yet, many current algorithmic fairness methodologies fail to account for the socially constructed nature of race and gender, instead treating these categories as fixed, indisputable, apolitical attributes.[42] These two types of demographic categories are highly contextual, and debates and legislation around gender and race classification are constantly evolving.

Misrepresentation can occur both when the categories used do not adequately represent individuals as they self-identify and when an individual is misclassified despite there being a representative category that they could have been classified as. To better understand the implications of misrepresentation, it's important to understand the different dimensions of identity and how these can lead to misrepresentation. With respect to racial identity, Roth[43] 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).[44] When racial data collection is conducted by observation, either by person or machine, there is the risk that an individual's observed race 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.[45]

Facial recognition technologies are a prominent case where the harm of misrepresentation occurs, since categorization is often based solely on observable characteristics. Additionally, many databases include a binary, physiological perspective of female and male, and consequently misrepresent individuals who do not self-identify with those categories.[46] Continuing to build databases that assume identity is a fixed, observable trait 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.[47]

DATA MISUSE AND USE BEYOND INFORMED CONSENT

Once collected, sensitive 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. Specifically 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 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). PATTERN was trained on data including demographic characteristics and criminal history for the purpose of assessing recidivism risk and providing guidance on recidivism reduction programming and productive activities for incarcerated people.[48] 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.[49] 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.

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 attribute data listed on people's social media profiles.[50] With this data, Clearview AI developed the world's most comprehensive facial recognition system, with a dataset beyond the scope of any government agency. Following public backlash, many of the social media platforms claimed that Clearview AI violated their policies. For example, LinkedIn sent Clearview AI a cease-and-desist letter

stating that scraping personal data was not permitted under their terms of service, and Facebook released a statement saying it demanded Clearview AI stop scraping data from its platforms in violation of company policy.[51] Individuals who have made profiles on these social media platforms and shared their images were not aware that their data was going to be used to develop a facial recognition system used by law enforcement agencies, nor were they asked for their consent.

Corporations collecting and using people's data to train and deploy ADMS face 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 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.[52] 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.[53] 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. 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 sensitive data consensually requires clear, specific, & limited use as well as strong security & protection following collection. Current consent practices are not meeting this standard.

Collecting sensitive data consensually requires clear, specific, and limited use as well as strong security and protection following collection. Current consent practices, including clickwraps and notice and consent frameworks, are not meeting this standard. Instead, these approaches overload individuals with descriptions and information that users see as boring and time-intensive.[54]

Risks to Communities

EXPANDING SURVEILLANCE INFRASTRUCTURE IN THE PURSUIT OF FAIRNESS

As discussed throughout this paper, there is often a trade-off between privacy and fairness when it comes to assessing discrimination and inequality. Zooming out from the scale of the individual to the scale of communities and groups of people, demographic data collection runs the risk of relying on, and justifying the expansion of, 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.[55]

Within this context of hypervisibility, 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. As such, it is likely that efforts at demographic data collection result in these groups being underrepresented in datasets. When this has happened in the past, well-meaning practitioners have sought to improve representation and system performance for these groups, motivating more targeted data collection efforts without much regard for the burdens and risks of this type of inclusion.[56]

In cases where there seems to be a tradeoff 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.[57]

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.[58] 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 system treatment across groups. At a high level, these risks center around essentializing or natural-

izing 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."[59]

The first of these risks, and certainly the one most frequently encountered and vocalized by practitioners,[60] 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.[61] It can also commonly occur in cases where demographic data is collected through inference or ascription by someone other than the data subject themselves. 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. 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.[62] In both 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[63] calls "administrative violence."

Demographic data collection efforts can reinforce oppressive norms & the delegitimization of disenfranchised groups.

Furthermore, data collected with overly narrow categories risks misrepresenting and obscuring subgroups subject to distinct forms of discrimination and inequality. This is often the case with Hmong communities which, despite facing extreme disenfranchisement, are seen as having opportunities typical of the larger AAPI category.[64]

Another way that categorization schema can be misaligned with various groups' experiences and lived realities is when the demographic variables themselves are too narrowly defined to capture all the dimensions of possible inequality. For example, as previously discussed, race can be self-identified (how an individual sees themselves), ascribed (how others see them), or relational (variable with regards to who or what someone is interacting with), and each of these dimensions carries with it different potential adverse treatments and effects.[65] 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.[66] 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.*

Finally, it is 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.[68] 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.[69] 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 it has been argued that race can (and often should be) seen as a "dynamic and interactive process, rather than a fixed thing that someone has."[70] 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.[71] Similar to the case with queer identities, attempts to come up with 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 their fluidity and mistreat them through other means, 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.[72]

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,[73] which in turn reinforces the notion that the differences between groups are natural and not a result of other social factors.[74] With regards to race, a categorization schema that

* AirBnB's Lighthouse project is a good example of this, as researchers accurately identified visually ascribed race as the dimension most likely for users to face discrimination for.[67] It is, however, still important to consider how ascription practices might differ between landlords, data labelers, and machine learning systems.

is conclusively not genetic or otherwise biological,[75] 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.[76] 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.[77]

Another way that risks can arise through relatively accurate categorization schema is through what philosopher Ian Hacking[78] refers to as a "looping effect" between categorization schema and a group's interaction with the world. 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. This effect is most often looked at in the context of psychiatric diagnostics, where individuals given a certain diagnosis start to adhere more closely to the diagnostic criteria over time, intentionally or not.[79] That being said, it is also applicable to other types of categories as well, such as gender and sexuality.[80] 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. 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 extremely important to consider when building systems that become increasingly tailored to users based on the information they provide, such as in the case of content recommendation algorithms as used by YouTube and TikTok.

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 we are asking already marginalized groups to share information for the purposes of assessing unfairness, it is imperative that the institution in question operationalizes fairness in a way that is aligned with these groups' interests and that we collect data that allows us to construct an accurate representation of the way members of these groups interact with systems.

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.[81] 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.[82] 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.[83] 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.[84] Despite this, commitments to neutrality remain the norm for many governmental and corporate policies.

Attempts to be neutral or objective often have the effect of reinforcing the status quo.

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 categorizing individuals from different groups at the same rate (i.e. true positive parity) 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[85] 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.[86] 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 and police and defendants and prosecutors 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.[87] Given that data collection efforts must be consciously designed, data always reflects one viewpoint or another about 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.[88] Historian Khalil Gibran Muhammad[89] has argued that the seemingly objective focus on data and statistical reasoning has replaced more explicitly racist understandings of racial difference, a shift made possible by the collection and analysis of 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

Risks in Algorithmic Decision-Making Systems

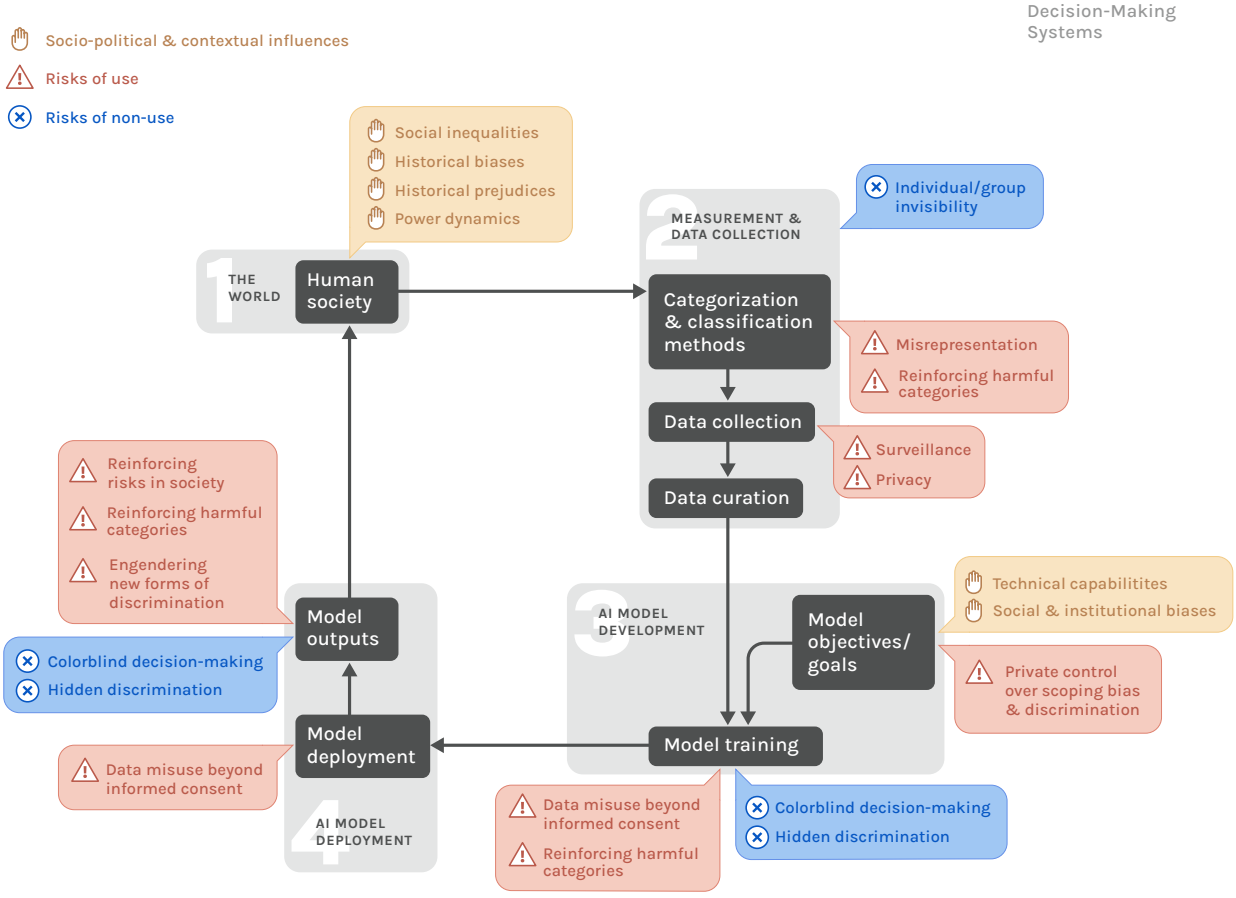

FIG. 1
Risks in Algorithmic Decision-Making Systems

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 that "each person should pay for his own risk," in the insurance industry.[90] 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 attributes statistically increase their risk.[91] To potentially mitigate some of this risk, institutions looking to collect demographic data should include more explicit documentation and commitments around what types of changes they are looking to make through assessing and bias and discrimination.

Conclusion

Balancing the risks of use and non-use of demographic data when it comes to fair algorithmic decision-making is ultimately a choice between risk trade-offs. In this paper we sought to provide an overview of some of the most pressing risks, but this is just the start of a much larger conversation. Each of these risks presents a whole suite of research questions that can only be tackled by individuals representing a diverse set of disciplines and industries. During 2022, Partnership on AI (PAI) will consult with partners to help develop a guidebook on how to responsibly collect and use demographic data to inform fair algorithmic decision-making. This guidebook will consider contexts in which it is appropriate to collect demographic data, assess what types of data are necessary, and provide recommendations on how organizations should collect and utilize sensitive information (including considerations around meaningful consent and compensation).

Additional research will seek to explore alternative data governance strategies, namely data cooperatives and data trusts. Open questions guiding our preliminary exploration into this area include: What factors should be considered for the establishment of a data collective? What type of third-party organization would be suitable for establishing and managing a data collective for sensitive data used to train machine learning systems?

If you have any feedback on this white paper or if you would like to receive updates about future demographic data research, please reach out to Sarah Villeneuve (sarah.v@partnershiponai.org) and McKane Andrus (mckane@partnershiponai.org).

Acknowledgements

We are grateful to the diverse set of stakeholders 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 are especially grateful to danah boyd (Microsoft Research, Data & Society), Nick Couldry (London School of Economics and Political Science), Emily Denton (Google Research), Ulises A. Mejias (SUNY Oswego), and Nithya Sambasivan (Google Research) for presenting at our CRAFT workshop on a number of the issues detailed in this report and helping us to start a wave of productive conversations.

Many PAI staff members contributed directly and indirectly to this work. In particular, McKane Andrus and Sarah Villeneuve, who led the project and the writing of the white paper, as well as Christine Custis, Tina Park, Hudson Hongo, Neil Uhl, and Penelope Bremner, who provided valuable ideas, advice, and assistance.

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.

To learn more about our Demographic Data Workstream, please visit our website, where you can fill out this form to become more involved with the demographic data community at PAI.

Notes

1 Andrus, M., Spitzer, E., Brown, J., & Xiang, A. (2021). "What We Can't Measure, We Can't Understand": Challenges to Demographic Data Procurement in the Pursuit of Fairness. *ArXiv:2011.02282* [Cs]. http://arxiv.org/abs/2011.02282

2 Andrus et al., 2021

3 Andrus et al., 2021

4 Tomasev, N., McKee, K. R., Kay, J., & Mohamed, S. (2021). Fairness for Unobserved Characteristics: Insights from Technological Impacts on Queer Communities. *ArXiv:2102.04257 [Cs]*. https://doi.org/10.1145/3461702.3462540

5 Andrus et al., 2021

6 Bakalar, C., Barreto, R., Bogen, M., Corbett-Davies, S., Hall, M., Kloumann, I., Lam, M., Candela, J. Q., Raghavan, M., Simons, J., Tannen, J., Tong, E., Vredenburgh, K., & Zhao, J. (2021). *Fairness On The Ground: Applying Algorithmic Fairness Approaches To Production Systems*. 12.

7 Balayn, A., & Gürses, S. (2021). *Beyond Debiasing*. European Digital Rights. https://edri.org/wp-content/uploads/2021/09/EDRi_Beyond-Debiasing-Report_Online.pdf

8 For a detailed discussion of the many kinds of data bias, see:

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., Kompatsiaris, I., Kinder-Kurlanda, K., Wagner, C., Karimi, F., Fernandez, M., Alani, H., Berendt, B., Kruegel, T., Heinze, C., … Staab, S. (2020). Bias in data-driven artificial intelligence systems—An introductory survey. *WIREs Data Mining and Knowledge Discovery*, *10*(3). https://doi.org/10.1002/widm.1356

Olteanu, A., Castillo, C., Diaz, F., & Kıcıman, E. (2019). Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. *Frontiers in Big Data*, *2*, 13. https://doi.org/10.3389/fdata.2019.00013

9 Rimfeld, K., & Malanchini, M. (2020, August 21). *The A-Level and GCSE scandal shows teachers should be trusted over exams results*. Inews.Co.Uk. https://inews.co.uk/opinion/a-level-gcse-results-trust-teachers-exams-592499

10 Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. *Proceedings of the International AAAI Conference on Web and Social Media*, *11*(1), 512–515.

Davidson, T., Bhattacharya, D., & Weber, I. (2019). Racial Bias in Hate Speech and Abusive Language Detection Datasets. *ArXiv:1905.12516 [Cs]*. http://arxiv.org/abs/1905.12516

11 Bogen, M., Rieke, A., & Ahmed, S. (2020). Awareness in Practice: Tensions in Access to Sensitive Attribute Data for Antidiscrimination. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 492–500. https://doi.org/10.1145/3351095.3372877

12 See, for example:

Executive Order on Advancing Racial Equity and Support for Underserved Communities Through the Federal Government. (2021, January 21). The White House. https://www.whitehouse.gov/briefing-room/presidential-actions/2021/01/20/executive-order-advancing-racial-equity-and-support-for-underserved-communities-through-the-federal-government/

Executive Order on Diversity, Equity, Inclusion, and Accessibility in the Federal Workforce. (2021, June 25). The White House. https://www.whitehouse.gov/briefing-room/presidential-actions/2021/06/25/executive-order-on-diversity-equity-inclusion-and-accessibility-in-the-federal-workforce/

13 Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual Fairness. Advances in *Neural Information Processing Systems, 30*. https://papers.nips.cc/paper/2017/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html

14 Harned, Z., & Wallach, H. (2019). Stretching human laws to apply to machines: The dangers of a'Colorblind'Computer. Florida State University Law Review, Forthcoming.

15 Washington, A. L. (2018). How to Argue with an Algorithm: Lessons from the COMPAS-ProPublica Debate. *Colorado Technology Law Journal*, *17*, 131.

16 Rodriguez, L. (2020). All Data Is Not Credit Data: Closing the Gap Between the Fair Housing Act and Algorithmic Decisionmaking in the Lending Industry. *Columbia Law Review*, *120*(7), 1843–1884.

17 Hu, L. (2021, February 22). *Law, Liberation, and Causal Inference*. LPE Project. https://lpeproject.org/blog/law-liberation-and-causal-inference/

18 See, for example:

Bonilla-Silva, E. (2010). *Racism Without Racists: Color-blind Racism and the Persistence of Racial Inequality in the United States*. Rowman & Littlefield.

Plaut, V. C., Thomas, K. M., Hurd, K., & Romano, C. A. (2018). Do Color Blindness and Multiculturalism Remedy or Foster Discrimination and Racism? *Current Directions in Psychological Science*, *27*(3), 200–206. https://doi.org/10.1177/0963721418766068

19 Eubanks, V. (2017). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.

20 Banco, E., & Tahir, D. (2021, March 9). *CDC under scrutiny after struggling to report Covid race, ethnicity data*. POLITICO. https://www.politico.com/news/2021/03/09/hhs-cdc-covid-race-data-474554

21 Banco & Tahir, 2021

22 Elliott, M. N., Morrison, P. A., Fremont, A., McCaffrey, D. F., Pantoja, P., & Lurie, N. (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), 69.

23 Shimkhada, R., Scheitler, A. J., & Ponce, N. A. (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), 81–102. https://doi.org/10.1007/s11113-020-09634-3

24 Poon, O. A., Dizon, J. P. M., & Squire, D. (2017). Count Me In!: Ethnic Data Disaggregation Advocacy, Racial Mattering, and Lessons for Racial Justice Coalitions. *JCSCORE, 3*(1), 91–124. https://doi.org/10.15763/issn.2642-2387.2017.3.1.91-124

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

26 Browne, S. (2015). Dark Matters: On the Surveillance of Blackness. In *Dark Matters*. Duke University Press. https://doi.org/10.1515/9780822375302

Eubanks, 2017

27 Farrand, T., Mireshghallah, F., Singh, S., & Trask, A. (2020). Neither Private Nor Fair: Impact of Data Imbalance on Utility and Fairness in Differential Privacy. *Proceedings of the 2020 Workshop on Privacy-Preserving Machine Learning in Practice*, 15–19. https://doi.org/10.1145/3411501.3419419

Jagielski, M., Kearns, M., Mao, J., Oprea, A., Roth, A., Sharifi-Malvajerdi, S., & Ullman, J. (2019). Differentially Private Fair Learning. *Proceedings of the 36th International Conference on Machine Learning*, 3000–3008. https://proceedings.mlr.press/v97/jagielski19a.html

Kuppam, S., Mckenna, R., Pujol, D., Hay, M., Machanavajjhala, A., & Miklau, G. (2020). Fair Decision Making using Privacy-Protected Data. *ArXiv:1905.12744* [Cs]. http://arxiv.org/abs/1905.12744

28 Quillian, L., Pager, D., Hexel, O., & Midtbøen, A. H. (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), 10870–10875. https://doi.org/10.1073/pnas.1706255114

Quillian, L., Lee, J. J., & Oliver, M. (2020). Evidence from Field Experiments in Hiring Shows Substantial Additional Racial Discrimination after the Callback. *Social Forces, 99*(2), 732–759. https://doi.org/10.1093/sf/soaa026

29 Cabañas, J. G., Cuevas, Á., Arrate, A., & Cuevas, R. (2021). Does Facebook use sensitive data for advertising purposes? *Communications of the ACM, 64*(1), 62–69. https://doi.org/10.1145/3426361

Datta, A., Tschantz, M. C., & Datta, A. (2015). Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination. *Proceedings on Privacy Enhancing Technologies, 2015*(1), 92–112. https://doi.org/10.1515/popets-2015-0007

30 Hupperich, T., Tatang, D., Wilkop, N., & Holz, T. (2018). An Empirical Study on Online Price Differentiation. *Proceedings of the Eighth ACM Conference on Data and Application Security and Privacy*, 76–83. https://doi.org/10.1145/3176258.3176338

Mikians, J., Gyarmati, L., Erramilli, V., & Laoutaris, N. (2013). Crowd-assisted search for price discrimination in e-commerce: First results. *Proceedings of the Ninth ACM Conference on Emerging Networking Experiments and Technologies*, 1–6. https://doi.org/10.1145/2535372.2535415

31 Cabañas et al., 2021

32 Leetaru, K. (2018, July 20). *Facebook As The Ultimate Government Surveillance Tool?* Forbes. https://www.forbes.com/sites/kalevleetaru/2018/07/20/facebook-as-the-ultimate-government-surveillance-tool/

Rozenshtein, A. Z. (2018). *Surveillance Intermediaries* (SSRN Scholarly Paper ID 2935321). Social Science Research Network. https://papers.ssrn.com/abstract=2935321

33 Rocher, L., Hendrickx, J. M., & de Montjoye, Y.-A. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications, 10*(1), 3069. https://doi.org/10.1038/s41467-019-10933-3

34 Cummings, R., Gupta, V., Kimpara, D., & Morgenstern, J. (2019). On the Compatibility of Privacy and Fairness. *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization – UMAP'19 Adjunct*, 309–315. https://doi.org/10.1145/3314183.3323847

Kuppam et al., 2020

35 Mavriki, P., & Karyda, M. (2019). Automated data-driven profiling: Threats for group privacy. *Information & Computer Security, 28*(2), 183–197. https://doi.org/10.1108/ICS-04-2019-0048

36 Angwin, J., & Parris, T. (2016, October 28). *Facebook Lets Advertisers Exclude Users by Race.* ProPublica. https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race

37 Barocas, S., & Levy, K. (2019). *Privacy Dependencies* (SSRN Scholarly Paper ID 3447384). Social Science Research Network. https://papers.ssrn.com/abstract=3447384

38 Bivens, R. (2017). The gender binary will not be deprogrammed: Ten years of coding gender on Facebook. *New Media & Society*, 19(6), 880–898. https://doi.org/10.1177/1461444815621527

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

40 Taylor, 2021

41 Draper and Turow, 2019

42 Hanna, A., Denton, E., Smart, A., & Smith-Loud, J. (2020). Towards a Critical Race Methodology in Algorithmic Fairness. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 501–512. https://doi.org/10.1145/3351095.3372826

Keyes, O., Hitzig, Z., & Blell, M. (2021). Truth from the machine: Artificial intelligence and the materialization of identity. *Interdisciplinary Science Reviews*, *46*(1–2), 158–175. https://doi.org/10.1080/03080188.2020.1840224

Scheuerman, M. K., Wade, K., Lustig, C., & Brubaker, J. R. (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), 1–35. https://doi.org/10.1145/3392866

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

44 Hanna et al., 2020

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

Keyes, O. (2019, April 8). *Counting the Countless*. Real Life. https://reallifemag.com/counting-the-countless/

Keyes et al., 2021

46 Scheuerman et al., 2020

47 Scheuerman et al., 2020

Stark, L., & Hutson, J. (2021). *Physiognomic Artificial Intelligence* (SSRN Scholarly Paper ID 3927300). Social Science Research Network. https://doi.org/10.2139/ssrn.3927300

48 U.S. Department of Justice. (2019). *The First Step Act of 2018: Risk and Needs Assessment System*. Office of the Attorney General.

49 Partnership on AI. (2020). *Algorithmic Risk Assessment and COVID-19: Why PATTERN Should Not Be Used*. Partnership on AI. http://partnershiponai.org/wp-content/uploads/2021/07/Why-PATTERN-Should-Not-Be-Used.pdf

50 Hill, K. (2020, January 18). *The Secretive Company That Might End Privacy as We Know It*. The New York Times. https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html

51 Porter, J. (2020, February 6). *Facebook and LinkedIn are latest to demand Clearview stop scraping images for facial recognition tech*. The Verge. https://www.theverge.com/2020/2/6/21126063/facebook-clearview-ai-image-scraping-facial-recognition-database-terms-of-service-twitter-youtube

52 *Regulation (EU) 2016/679 (General Data Protection Regulation)*, (2016) (testimony of European Parliament and Council of European Union). https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679

53 Obar, J. A. (2020). Sunlight alone is not a disinfectant: Consent and the futility of opening Big Data black boxes (without assistance). *Big Data & Society*, *7*(1), 2053951720935615. https://doi.org/10.1177/2053951720935615

54 Obar, 2020

Oeldorf-Hirsch, A., & Obar, J. A. (2019). Overwhelming, Important, Irrelevant: Terms of Service and Privacy Policy Reading among Older Adults. *Proceedings of the 10th International Conference on Social Media and Society*, 166–173. https://doi.org/10.1145/3328529.3328557

55 Benjamin, R. (2019). *Race after technology: Abolitionist tools for the new Jim code*. Polity.

Browne, 2015; Eubanks, 2017

56 Hoffmann, 2020

57 Rainie, S. C., Kukutai, T., Walter, M., Figueroa-Rodríguez, O. L., Walker, J., & Axelsson, P. (2019). *Indigenous data sovereignty*.

Ricaurte, P. (2019). Data Epistemologies, Coloniality of Power, and Resistance. *Television & New Media*, 16.

Walter, M. (2020, October 7). *Delivering Indigenous Data Sovereignty*. https://www.youtube.com/watch?v=NCsCZJ8ugPA

58 See, for example:

Bowker, G. C., & Star, S. L. (1999). *Sorting things out: Classification and its consequences*. MIT Press.

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

Hacking, I. (1995). The looping effects of human kinds. In *Causal cognition: A multidisciplinary debate* (pp. 351–394). Clarendon Press/Oxford University Press.

Hanna et al., 2020

Hu, L., & Kohler-Hausmann, I. (2020). *What's Sex Got to Do With Fair Machine Learning?* 11.

Keyes (2019)

Zuberi, T., & Bonilla-Silva, E. (2008). *White Logic, White Methods: Racism and Methodology*. Rowman & Littlefield Publishers.

59 Hanna et al., 2020

60 Andrus et al., 2021

61 Bivens, 2017

62 Hamidi, F., Scheuerman, M. K., & Branham, S. M. (2018). Gender Recognition or Gender Reductionism?: The Social Implications of Embedded Gender Recognition Systems. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems – CHI '18*, 1–13. https://doi.org/10.1145/3173574.3173582

Keyes, 2018

Keyes et al., 2021

63 Spade, D. (2015). Normal Life: Administrative Violence, Critical Trans Politics, and the Limits of Law. In *Normal Life*. Duke University Press. https://doi.org/10.1515/9780822374794

64 Fu, S., & King, K. (2021). Data disaggregation and its discontents: Discourses of civil rights, efficiency and ethnic registry. *Discourse: Studies in the Cultural Politics of Education, 42*(2), 199–214. https://doi.org/10.1080/01596306.2019.1602507

Poon et al., 2017

65 Hanna et al., 2020

66 Saperstein, A. (2012). Capturing complexity in the United States: Which aspects of race matter and when? *Ethnic and Racial Studies*, *35*(8), 1484–1502. https://doi.org/10.1080/01419870.2011.607504

67 Basu, S., Berman, R., Bloomston, A., Cambell, J., Diaz, A., Era, N., Evans, B., Palkar, S., & Wharton, S. (2020). *Measuring discrepancies in Airbnb guest acceptance rates using anonymized demographic data*. AirBnB. https://news.airbnb.com/wp-content/uploads/sites/4/2020/06/Project-Lighthouse-Airbnb-2020-06-12.pdf

68 Keyes, 2019

Ruberg, B., & Ruelos, S. (2020). Data for queer lives: How LGBTQ gender and sexuality identities challenge norms of demographics. *Big Data & Society, 7*(1), 2053951720933286. https://doi.org/10.1177/2053951720933286

69 Tomasev et al., 2021

70 Pauker, K., Meyers, C., Sanchez, D. T., Gaither, S. E., & Young, D. M. (2018). A review of multiracial malleability: Identity, categorization, and shifting racial attitudes. *Social and Personality Psychology Compass*, *12*(6), e12392. https://doi.org/10.1111/spc3.12392

71 Pauker et al., 2018

72 Ruberg & Ruelos, 2020

73 Braun, L., Fausto-Sterling, A., Fullwiley, D., Hammonds, E. M., Nelson, A., Quivers, W., Reverby, S. M., & Shields, A. E. (2007). Racial Categories in Medical Practice: How Useful Are They? *PLOS Medicine, 4*(9), e271. https://doi.org/10.1371/journal.pmed.0040271

74 Hanna et al., 2020

75 Morning, A. (2014). Does Genomics Challenge the Social Construction of Race?: *Sociological Theory*. https://doi.org/10.1177/0735275114550881

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

77 Barabas, 2019

78 Hacking, 1995

79 Hacking, 1995

80 Dembroff, 2018

81 Andrus et al., 2021

Holstein, K., Vaughan, J. W., Daumé III, H., Dudík, M., & Wallach, H. (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*, 1–16. https://doi.org/10.1145/3290605.3300830

Rakova, B., Yang, J., Cramer, H., & Chowdhury, R. (2021). Where Responsible AI meets Reality: Practitioner Perspectives on Enablers for shifting Organizational Practices. *ArXiv:2006.12358 [Cs]*. https://doi.org/10.1145/3449081

82 Wachter, S., Mittelstadt, B., & Russell, C. (2021). Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI. *Computer Law & Security Review, 41*. https://doi.org/10.2139/ssrn.3547922

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

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

83 Balayn & Gürses, 2021

84 Fazelpour, S., & Lipton, Z. C. (2020). Algorithmic Fairness from a Non-ideal Perspective. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 57–63. https://doi.org/10.1145/3375627.3375828

Green, B., & Viljoen, S. (2020). Algorithmic realism: Expanding the boundaries of algorithmic thought. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 19–31. https://doi.org/10.1145/3351095.3372840

85 Green & Viljoen, 2020

86 Gitelman, L. (2013). *Raw Data Is an Oxymoron*. MIT Press.

87 Barabas, C., Doyle, C., Rubinovitz, J., & Dinakar, K. (2020). *Studying Up: Reorienting the study of algorithmic fairness around issues of power*. 10.

88 Crooks, R., & Currie, M. (2021). Numbers will not save us: Agonistic data practices. *The Information Society, 0*(0), 1–19. https://doi.org/10.1080/01972243.2021.1920081

89 Muhammad, K. G. (2019). *The Condemnation of Blackness: Race, Crime, and the Making of Modern Urban America, With a New Preface*. Harvard University Press.

90 Ochigame, R., Barabas, C., Dinakar, K., Virza, M., & Ito, J. (2018). Beyond Legitimation: Rethinking Fairness, Interpretability, and Accuracy in Machine Learning. *International Conference on Machine Learning*, 6.

91 Ochigame et al., 2018