Towards a Critical Race Methodology in Algorithmic Fairness

Alex Hanna*
Emily Denton*
Andrew Smart
Jamila Smith-Loud
{alexhanna,dentone,andrewsmart,jsmithloud}@google.com

ABSTRACT

We examine the way race and racial categories are adopted in algorithmic fairness frameworks. Current methodologies fail to adequately account for the socially constructed nature of race, instead adopting a conceptualization of race as a fixed attribute. Treating race as an attribute, rather than a structural, institutional, and relational phenomenon, can serve to minimize the structural aspects of algorithmic unfairness. In this work, we focus on the history of racial categories and turn to critical race theory and sociological work on race and ethnicity to ground conceptualizations of race for fairness research, drawing on lessons from public health, biomedical research, and social survey research. We argue that algorithmic fairness researchers need to take into account the multidimensionality of race, take seriously the processes of conceptualizing and operationalizing race, focus on social processes which produce racial inequality, and consider perspectives of those most affected by sociotechnical systems.

CCS CONCEPTS

• **Applied computing** → **Sociology**; • **Social and professional topics** → *Race and ethnicity*.

KEYWORDS

algorithmic fairness, critical race theory, race and ethnicity

ACM Reference Format:
Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a Critical Race Methodology in Algorithmic Fairness. In *Conference on Fairness, Accountability, and Transparency (FAT* '20), January 27–30, 2020, Barcelona, Spain.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3351095.3372826

> The problem does not end with the collection of racial data; it only begins. The problem accelerates when we attempt to analyze these data statistically... The racialization of data is an artifact of both the struggles to preserve and to destroy racial stratification. Before the data can be deracialized, we must deracialize the social circumstances that have created racial stratification.
> – Tufuku Zuberi [125, pp. 102]

1 INTRODUCTION

In recent years, there has been increasing recognition of the potential for algorithmic systems to reproduce or amplify existing social inequities. In response, the research field of algorithmic fairness has emerged. This rapidly evolving research area is focused on developing tools and techniques with aspirations to make development, use, and resulting impact of algorithmic systems conform to various social and legal notions of fairness. The concept of fairness, in addition to being situational, evolving, and contested from a number of philosophical and legal traditions, can only be understood in reference to the different social groups that constitute the organization of society. Consequently, the vast majority of algorithmic fairness frameworks are specified with reference to these social groups, often requiring a formal encoding of the groups into the dataset and/or algorithm.

However, most social groups relevant to fairness analysis reflect highly contextual and unstable social constructs. These social groups are often defined with recourse to legal anti-discrimination concepts such as "protected classes," which, in the US, refers to race, color, national origin, religion, sex, age, or disability. However, the process of drawing boundaries around distinct social groups for fairness research is fraught; the construction of categories has a long history of political struggle and legal argumentation.

Numerous recent works have highlighted the limitations of current algorithmic fairness frameworks [53, 108]. Several of these critiques point to the tendency to abstract away critical social and historical contexts and minimize the structural conditions that underpin problems of algorithmic unfairness. We build on this critical research, focusing specifically on the use of race and racial categories within this field. IN this literature, the topic of the instability of racial categories has gone relative unexplored, with the notable exception of Benthall and Haynes [12], which we discuss in detail below.

Race is a major axis around which algorithmic allocation of resources and representation is bound. It may indeed be the most significant axis, given attention by investigative journalists (e.g. [7]) and critical race and technology scholars (e.g. [11, 20, 21, 24, 85]).

[*] Both authors contributed equally to this research.

FAT '20, January 27–30, 2020, Barcelona, Spain*

ACM ISBN 978-1-4503-6936-7/20/02.
https://doi.org/10.1145/3351095.3372826

Because of this, it is imperative that the race-based methodologies and racial categories themselves are interrogated and critically evaluated.

In this paper we develop several lines of critique directed at the treatment of race within algorithmic fairness methodologies. In short, we observe that current methodologies fail to adequately account for the socially constructed nature of race, instead adopting a conceptualization of race as a fixed attribute. This manifests in the widespread use of racial categories as if they represent natural and objective differences between groups. Treating race as an attribute, rather than a structural, institutional, and relational phenomenon, in turn serves to minimize the structural aspects of algorithmic unfairness.

The process of operationalizing race is fundamentally a project of racial classification and thus must be understood as a political project, or, more specifically, what Omi and Winant [86] refer to as a racial project. The tools for measuring individual characteristics, such as national censuses and other population-level survey instruments, were and are still often based in the politics of racial oppression and domination. While we acknowledge the importance of measuring race for the purposes of understanding patterns of differential performance or differentially adverse impact of algorithmic systems, in this work, we emphasize that data collection and annotation efforts must be grounded in the social and historical contexts of racial classification and racial category formation. To oversimplify is to do violence, or even more, to re-inscribe violence on communities that already experience structural violence.

It is useful to distinguish between two ways in which race comes into play in algorithmic fairness research: (i) the operationalization of race, i.e. the process of converting the abstract concept of race into something that is concrete and measurable and (ii) the use of racial variables within algorithmic frameworks. These aspects are tightly interconnected since race must be operationalized before racial variables can be utilized.

Our contributions are as follows: we review the use of race in prior fairness work and discuss an intervention in the debate on racial categories. We then survey the history of racial classification, scientific racism, and racial classification in national censuses. We introduce the notion of multidimensionality of race, and subsequently review how different disciplines have dealt with the tricky problem of operationalizing race, most notably in biomedical, survey, and public health research. In light of these discussions, we address how race has been treated in the group fairness and disaggregated analysis paradigms. Lastly, we argue that fairness researchers need to take into account the multidimensionality of race, take seriously the processes of conceptualizing and operationalizing race, focus on processes of racism, and consider perspectives of those most affected by sociotechnical systems.

2 THE PROBLEM WITH RACIAL CATEGORIES

In perhaps the best-known debate surrounding algorithmic fairness, Julia Angwin and her colleagues illustrated the differing rates at which Black defendants had been issued high pre-trial risk assessment scores by the COMPAS recidivism prediction algorithm compared to white defendants [7]. Equivant – then, Northpointe, the company which developed COMPAS – defended the system by reanalyzing the ProPublica data [28] and ProPublica later responded [6]. The debate has become a cornerstone of algorithmic fairness research; to date, the original story has some 700 citations on Google Scholar.

The Propublica analysis of COMPAS, as well as the responses from Northpointe and other secondary analyses performed by third-party researchers, relied on data from the Broward County Sheriff's Office which was obtained through a public records request. Race classifications in this data identified defendants as Black, White, Hispanic, Asian, Native American or Other. These categories look familiar – they are nearly identical to the way that the US Census defines race. However, there are the notable absences of the categories "Native Hawaiian or Other Pacific Islander", and there is a redefinition of "Hispanic" as a race rather than an ethnicity.

In the methodological appendix to the original article [69], Jeff Larson and other ProPublica authors do not delve into how the race of the defendant is measured. Larson admits that he did not know how Broward County classified individuals [1]. It's not clear why Broward County used the modified Census racial schema, but, as detailed below, it may have to do with a 1977 directive from the federal US Office of Management and Budget. In general, approaches to measuring race for police records are inconsistent across municipalities. They can rely on self-identification, state records, or observation by criminal justice workers. In her work on racial disparities in incarceration in Wisconsin, sociologist Pamela Oliver notes that the race of a single individual can change across records even within a single jurisdiction [2]. Therefore, even in the most famous debate of the field, we don't know why the data take on a particular racial schema, nor do we have information about how defendants are racially categorized.

2.1 Using race-like categories?

While an area of major concern for critical race and technology scholarship, the use of racial categories in algorithmic fairness research (i.e. the research community which has emerged around venues like FAT* and AIES) has largely gone unquestioned. There are two important exceptions. First, in the original Gender Shades paper which helped establish intersectional testing, Buolamwini and Gebru note the instability of race and ethnicity labels, which justifies their use of skin tone [22]. Second, the critique levied by computational social scientist Sebastian Benthall and critical race scholar Bruce D. Haynes [12] describes how little attention has been given to the meaning of protected class labels, the political orientation of class labels, and the normative assumptions of machine learning design. They focus on the typical referent group – African-Americans in the US – and discuss how "racial classification is embedded in state institutions, and reinforced in civil society in ways that are relevant to the design of machine learning systems." Race is invoked because "racial classification signifies social, economic, and political inequities anchored in state and civic institutional practices."

We wholly agree with their overall argument, which is that algorithmic fairness research has been largely silent on the issues

[1] Personal correspondence, Jeff Larson, August 19, 2019.

[2] Personal correspondence, Pamela Oliver, August 19, 2019.

of racial statistics and categories, and that in that silence, has thus reified them. In their criticism of the COMPAS debate, we echo their argument:

> rather than taking racial statistics... at face value, the process that generates them and the process through which they are interpreted should be analyzed with the same rigor and skepticism as the recidivism prediction algorithm. Thematically, we argue that racial bias is far more likely to come from human judgments in data generation and interpretation than from an algorithmic model, and that this has broad implications for fairness in machine learning [12, pp. 291].

We diverge from their critique, not necessarily in their problem formation, but in their solution. They offer up a narrow path to follow which plots a course between the Scylla of "[m]achines that ... [reify] racial categories that are inherent unfair" and the Charybdis of "systems that allocate resources in ways that are blind to race [which] will reproduce racial inequality in society" [12, pp. 294]. In their solution, they propose using an unsupervised machine learning method to create race-*like* categories which aim to address "historical racial segregation with reproducing the political construction of racial categories." One virtue of this method is that it is attentive to creating a metric which makes sense with respect to a particular social inequality, that is, spatial segregation. This is something we agree with in our adoption of a multidimensional perspective below.

We'd like to raise a few complexities in their formulation which make us hesitant to adopt their solution. First, it would be a grave error to supplant the existing categories of race with race-like categories inferred by unsupervised learning methods. Despite the risk of reifying the socially constructed idea called race, race does exist in the world, as a way of mental sorting, as a discourse which is adopted, as a social thing which has both structural and ideological components. In other words, although race is social constructed, race still has power. To supplant race with race-like categories for the purposes of measurement sidesteps the problem.

Second, supplanting race with race-like categories depends highly on context, namely how race operates within particular systems of inequality and domination. Benthall and Haynes restrict their analysis to that of spatial segregation, which is to be sure, an important and active research area and subject of significant policy discussion (e.g. [76, 99]). However, that metric may appear illegible to analyses pertaining to other racialized institutions, such as the criminal justice system, education, or employment (although one can readily see their connections and interdependencies). The way that race matters or pertains to particular types of structural inequality depends on that context and requires its own modes of operationalization.

Third, and relatedly, as Wendy Chun and Ruha Benjamin have discussed [11, 24], race operates both *with* and *as* technology. At the same time we focus on the ontological aspects of race (what is race, how is it constituted and imagined in the world), it is necessary to pay attention to what we do with race and measures which may be interpreted as race. The creation of metrics and indicators which are race-like will still be interpreted as race. As we discuss more below, the example of genomics research is indicative. Even as genomics researchers warn not to interpret genetic ancestry as race or ethnicity [15, 122], this has not prevented people (e.g. customers of direct-to-consumer genetic testing companies) from interpreting them as such [98].

Finally, we'd like to underline the *infrastructural* criticism embedded in the act of categorization itself. By infrastructure, we refer to the way that classifications and standards form the infrastructure of our existing information society [17]. Inverting infrastructures and seeking their fissures allows us to understand how they are used in practice and how politics are embedded in their creation. This point dovetails with the critiques levied by abolitionist activists, scholars, and technologists [1, 9, 11]. J. Khadijah Abdurahman [1] argues that to study algorithmic fairness is to sidestep the problem beyond the algorithmic frame, that we, with communities who are disproportionally affected by redlining, predictive policing, and surveillance, should not just contest how "protected classes within algorithms are generated-but [should] viscerally reject the notion human beings should be placed in cages in the first place." The task of using racial classifications at all requires us to expand the frame to consider what the larger implications of classifying are. Who is doing the classifying? For what purpose are they classifying and to what end?

3 HISTORIES OF RACIAL CATEGORIZATION

We recount histories of racial category construction by administrative and scientific bodies as a means of highlighting the decidedly social constructivist notion of race. As Hacking [48] has said, social constructivism is the act of "making up people." Social construction does not mean that things in the world are not "real" [31], but that that thing – be it a racial, gender, or class category – was brought into existence or shaped by historical events, social forces, political power, and/or colonial conquest, all of which could have been very different. When we demonstrate that something is socially constructed, it becomes clear that it could be constructed differently, and then we can start to demand changes in it [48, pp. 6-7].

Race as a concept is widely acknowledged by the social sciences to be a social construction. In the racial formation framework, Omi and Winant [86] discuss how race is both real and socially constructed. The social constructedness of race decenters the concept as a property of individuals determined by phenotypical properties. Instead, social constructivism places race within specific history and context. The constructedness of race in any given point in time is tied to the specific *racial project*. A racial project is "simultaneously an interpretation, representation, or explanation of racial identities and meanings, and an effort to organize and distribute resources (economic, political, cultural) along particular racial lines" [86, pp. 56]. Those racial projects can range in scale and kind, in the microinteractional to the structural.

Race is not something that arrived whole cloth, but needed to be made up. Race, although socially constructed, of course has very salient material effects. Accordingly, the most modern understanding of race argues that it is inauthentic to conceptualize race as a natural property of individuals. The "naturalization" of race and racial categories do not come from nowhere. It has been constructed and been continually reproduced as part of a project of upholding a particular type of racial project – one that is historically and culturally bound. Historian Ibram X. Kendi, for instance, thoroughly outlines the history of race and racial projects from 15-century Europe to the present [64]. Projects that "misrecognize" race as natural [27] are driven by the hegemonic nature of racial projects [86]. It would be more accurate to describe race as relational and as a property of institutions, organizations, and larger structures [93]. Race may be more accurately discussed as having relational qualities, with dimensions that are symbolic or based on phenotype, but that are also contingent on specific social and historical contexts.

3.1 Classification and the Racial History of Social Statistics

Classification is a process imbued with social, economic, and organizational imperatives [17, 38]. Those who do the categorizing have their own institutional and occupational priorities. Categories themselves become a type of infrastructure: they are the ground upon which other structural and ideational elements are built. When they work, they are invisible, but when they break down, the boundaries and assumptions begin to show. With the development of markets and the rise of technological innovation and actuarial science, classification of individuals has moved from the more general assessment based on subgroup distinction to individuation based on market imperatives.

Although in the modern neoliberal era, markets have been an emerging force in categorization and segmentation [39], the entity tasked with categorization above all others has been the state. Almost as soon as humans began living in agriculturally-based communities, it became necessary to count the number of people living under the control of a certain party [107]. Categorization was also used to streamline counting items of like type – such as trees in a forest, grain, or human beings. Further, since early grain-based agriculture necessitated vast numbers of essentially slave-laborers, rulers needed to know how many slaves there were and moreover, who counted as a slave. As Scott argues, during the agricultural revolution there was initially strong resistance to transitioning from life as a hunter-gatherer to a more hierarchical, regimented, and predictable existence as a slave harvesting grain and to "being counted" by newly-created states [107].

The modern discipline of statistics grew out of a bureaucratic need to manage large populations and natural resources [52]. More specifically and perniciously, the field of social statistics emerged from the need to make differentiations between white Europeans and their descendants, and other peoples. Francis Galton was a major figure in the development of social statistics; he was also a eugenicist and a major proponent of using statistics as a means to justify racial superiority of European-origin peoples [16, 124, 125]. As Galton writes in his *Hereditary Genius*:

> The natural ability of which this book mainly treats, is such as a modern European possesses in a much greater average share than men of the lower races. There is nothing either in the history of domestic animals or in that evolution to make us doubt that a race of sane men may be formed, who shall be as much superior mentally and morally to the modern European as the modern European is to the lowest of the Negro races (quoted in [124]).

The practice of racial classification that emerged in social statistics is closely tied to the project of nation-building, in multiple senses. In the first case, racial classification was a prerequisite to the project of European colonization of the Americas. The racialization of chattel slavery is a modern phenomenon. European colonial projects developed alongside Enlightenment ideas of liberal democracy. Creating structures of racial stratification necessitated ideational components that would legitimate the enslavement of Africans [125]. Until the early 20th century in the West and the rise of cultural anthropologists such as Franz Boas, there was widespread support for eugenics and scientific racism, and the belief that the state must intervene to protect the physical and mental characteristics of the white race. In 1926, 23 of 48 states had laws permitting sterilization [106]. At the height of miscegenation legislation (that is, laws against marriages which were "interracial" or "cross-cultural" in nature), 41 American colonies and states had laws against the practice [88]. As the project of nation-building (more specifically, colonization of non-European nations) waned in the 1960s, so did the popularity of eugenics as a scientific practice. However, as Syed Mustafa Ali notes, the scientific racism undergirding colonialism has persisted as "'sedimented' ways of knowing and being – based on systems of categorisation, classification, and taxonomisation and the ways that these are manifested in practices, artefacts and technologies" [2]. Technologies of racial classification, such as blood quantum for Native Americans, persist and upon legacies of settler colonialism. These technologies become enshrined in folk understandings of racial percentages, such as genetic ancestry [113], or in law, as blood quantum remains a legal requisite to tribal membership, tribal constitution, and land claims.

Second, since the founding of the Americas, delineating racial boundaries has been part and parcel of the state-building project. Assessing and evaluating racial and ethnic boundaries is a state practice that coincides not only with a cultural understanding of who belongs to the polity, but has and continually is a prerequisite for formal citizenship in many countries. In the pre-war United States, appeals to citizenship for non-native born residents were fundamentally appeals to whiteness. Lopez documents how appeals to whiteness and the boundaries of the category of whiteness flexed and contracted in legal argumentation [73]. The 1790 Naturalization Act restricted citizenship to "any alien, being a free white person" who had been in the United States for at least two years. It wasn't until the passage of the Immigration and Nationality Act of 1952 that the explicit racialized nature of naturalization was restricted, although the Act laid the groundwork for the current restrictive race-based ratio system which exists today. Maghbouleh demonstrates how, prior to 1952, Iranians could be used as a "racial hinge" to argue both for and against claims to whiteness, and how citizenship claims critically depended on who could make this claim [75].

Third, national censuses are a critical component of state machinery. State enumeration of populations works to facilitate the administrative and bureaucratic functions of the modern nation-state. As Scott points out:

> State simplifications such as maps, censuses, cadastral lists, and standard units of measurement represent techniques for grasping a large and complex reality; in order for officials to be able to comprehend aspects of the ensemble, that complex reality must be reduced to schematic categories. The only way to accomplish this is to reduce an infinite array of detail to a set of categories that will facilitate summary descriptions, comparisons, and aggregation. The invention, elaboration, and deployment of these abstractions represent, as Charles Tilly [116] has shown, an enormous leap in state capacity – a move from tribute and indirect rule to taxation and direct rule [106, pp. 77].

Furthermore, censuses not only quantify members of racial social groups, but also *create and constitute the boundaries of these social groups themselves.* Melissa Nobles outlines how racial classifications are taken for granted by policymakers, but how national censuses not only enumerate polity members into racial categories, but also help to reinscribe racial categories themselves, that is, "[c]ensus-taking is one of the institutional mechanisms by which racial boundaries are set" [112, pp. xi]. In the US, the national census is the sole basis for which representation is guaranteed in the House of Representatives and the Electoral College. In the antebellum US, the Three-Fifths Compromise, in which Black slaves were counted as three-fifths of a human for the purposes of appropriation and taxation, is the most dramatic example of racial classification in counting populations. Whereas the passage of the 14th Amendment guaranteed all Americans being counted as a whole person regardless of race, racial categories defined in the Census in the latter part of the 19th Century reflected fractional referents to Black racial lineage based in the ascendant movement of eugenics and race science mentioned above. Categories of "octoroon" and "quadroon", along with "mulatto" denoted fractional heredities of the Black population [112].

Because of the allocative nature of census-taking and the bureaucracies associated with the practice, what goes into the census is politically contested by mobilizing racial and ethnic groups. The political contestation of the census has been laid bare recently with the attempted introduction of a citizenship question into the US Census and its potential for its dampening effects on participation by Latinx groups [72]. For instance, after the 1950 Census, in light of the discoveries of dramatic undercounts of non-white populations and with rising pressures from the Civil Rights Movement, the 1960 Census shifting from interview identification to asking individuals to self-identify their race. As Snipp rightly points out, this procedural change was a "fundamental redefinition of race." Population statistics formerly based on phenotypical appearances were now based on "cultural affiliation and other deeply held personal considerations beyond the pale of conventional demographic inquiry" [112, pp. 570]. This shift was not only bureaucratic but also (perhaps unknowingly by the Census Bureau) ontological: race as measured by the Census aligned more with the social constructionism and less with essentialism. The Office of Management and Budget (OMB) Directive No. 15 of 1977 instituted a standard in racial classifications used by government agencies and developed a standard for five distinct groups: (a) American Indians and Alaska Natives, (b) Asian and Pacific Islanders, (c) Non-Hispanic Blacks, (d) Non-Hispanic Whites, and (e) Hispanics. The impact of this directive had far-reaching implications, as this categorization "permeated every level of government, many if not most large corporations, and many other institutions such as schools and nonprofit organizations" ([87], cited in [112]). The 1997 revision of Directive No. 15 separated out Native Hawaiians and other Pacific Islanders, set Hispanics and Latinos apart as an ethnic group (rather than a racial group), and allowed respondents to provide multiple responses to racial heritage [112].

Dramatic changes in governance have large downstream effects for the accuracy of census counts and their categories. Khalfani et al. document the differences between the South African numbers before and after the reign of Mandela's African National Congress, and find that the raw counts and the estimations from aerial photography both dramatically undercount the majority African populations, as well as women [65]. The US Census Bureau spends a nontrivial amount of effort getting enumeration correct for marginalized groups; however, those efforts can be stymied with the winds of partisanship and political administration. Even without the issues raised by a potential citizenship question, decreased or flat-lined funding has left the Census Bureau understaffed and unable to thoroughly test new methodologies [32].

In sum, the act of classification cannot be divorced from the group who is doing the classification and the organizational and institutional pressures of performing classification. Racial classification has a long history within quantitative social science, and, in some ways, is the reason that social statistics have developed the way they did. That history is grounded in eugenicist thought and practice, and much of that work provided the infrastructure for census-taking in the US, Canada, Latin America, and South Africa. More expansive categorization resulted from social and political movements in those countries. In short, categorization itself is a technological infrastructure within which institutional racism continues to reside. In this way, we can realize the ways in which race intersects with technology but also as a technology in and of itself. The framing of race as technology shifts the terrain of the discussion – rather, from what race is and how we can classify individuals, to what we can do with race and how to make it do different (and explicitly anti-racist) things.

3.2 The multidimensionality of race

As highlighted by the 1960 Census shift from ascribed to self-identification, classification is largely contingent on the particular appraisal of race used. When we say "race", we may be discussing self-identification, but we also may be referring to phenotypical features or observed assessments from third parties. Racial classifications are uneasily balanced not only on the particular unstable equilibrium of racial projects, but also on the micro-level processes of race appraisals themselves. When it comes to measurement and operationalization, "race" is not a single variable, but many differing and sometimes competing variables. Roth [95] refers to this constellation of variables as the "multiple dimensions" of race. Race manifests in many ways in the real world. Each of these different dimensions can be measured in a different manner for empirical research, and each of these different dimensions have distinct downstream outcomes. It may therefore be more appropriate to study a particular outcome using measurement which reflects a respective dimension of race. Table 1 reproduces, with minor amendments, the table from [95] which outlines each of these dimensions.

Racial identity and *racial self-classification* refer to subjective self-identification of race. In theory, self-identification is one dimension but in practice these two are measured in different ways and have differing downstream outcomes. Racial identity can be thought of as the sense of self, the identity and broad group with which someone belongs. Accordingly, its measurement instrument is that of an open-ended interviewer or survey question. Racial self-classification refers to the discrete categories that one marks on intake forms, census surveys, and self-identification documents for employment. This dimension, as noted earlier, is highly constrained by the categories determined *a* priori by institutional and organizational directives, such as OMB Directive No. 15.

Observed race refers to the race which others believe ascribed to an individual. This may be based upon first impression by an interviewer or annotator. The third-party can make an assessment based solely on *appearance* or *interaction*. Appearance-based observed race depends on readily observable characteristics, whereas interaction-based observed race depends on language, accent, and other physical, aural, or social cues. Reflected race is the race which an individual believes that others ascribe to them. Although this is a first-person evaluation, it is the first-person evaluation of a third-party appraisal.

Phenotype refers to the set of objective characteristics which characterize appearance. A significant literature revolves around skin color appraisals as a means for determining particular social outcomes (e.g. [82, 114]) and the use of skin color as a means of evaluating facial recognition systems (e.g. [22, 92]). Other race-related phenotypical features include hair texture and color, eye shape and color, and lip shape. The related outcomes for phenotypical features are the same as those for observed race, but are not constrained in measurement by discrete racial categorization[3].

The dimensions of race noted here are, of course, unstable across time, place, and context. Just as racial classifications themselves are tied to particular racial projects, modes of self-identification, observed race, and reflected race will be tied to the dynamics of that project. Self-identification has been shown to be responsive to one's arrest records [89] and to one's class position [115]. Observed race depends on the interviewer or annotator who is doing the observing. Reflected race depends not only on phenotype but also on the social status of the group in which the individual believes they are being recognized as. These positions shift across time, not only across the life course but also across the *longue durée* of different racial projects.

Three major implications follow from the of multiple dimensions paradigm. The first is that there may be inconsistencies across different dimensions of race. A single individual can have differing racial appraisals. In the health field, for instance, Roth cites several studies in which individual self-identification differs from medical records, interviewer classifications, or death certificates. Second, the dimensions can cross the boundaries between each other. Phenotype and reflected race can impact self-identification but do not fully determine it. However, the third and most important implication, also highlighted in Table 1, is that different dimensions will be associated with differing outcomes. Therefore, using a measurement of race which does not reflect social processes associated with it may misrepresent the scope of racial disparities at best and underreport them at worst.

4 LESSONS FROM OTHER DISCIPLINES

The variance in what the concept of race can mean has been noted as "antithetical to the tenets of scientific research, which, in its ideal form, demands that analytical variables be consistent and their categories mutually exclusive" [70]. Yet, the social centrality of race positions it as a critical concept in many fields of study. In this section, we review how other disciplines, such as public health research, biomedical research, and studies of social inequality, have grappled with the use of racial categories within their disciplines. We surface these discussions to understand how algorithmic fairness research can learn from these methodological approaches.

4.1 Limitations in operationalizing race

Scholars from a range of disciplines have observed a widespread lack of clarity around the use of racial variables within their respective fields. Race is often inconsistently conceptualized and measured across studies; definitions and processes of operationalization tend to be insufficiently documented. These inconsistencies pose serious challenges for the validity and utility of research results. For example, inconsistent categories and classification schemes can result in mismatches between different measures of race for the same individual across different time points or across studies. This in turn can significantly impact health statistics [62, 101], population statistics [8, 74], and measures of social inequality [102]. Country-specific understandings of what "race" references and local racial categorization schemes pose significant challenges for international comparisons [96].

Standardizing racial taxonomies and measurement practices does not sufficiently mitigate the many concerns scholars have raised regarding the use of racial categories in scientific studies. In fact, the widespread uncritical adoption of racial categories, standardized or not, can erode awareness of the social and political histories of racial taxonomies and reify racial categories as natural kinds [19, 30, 42, 43, 78, 100, 111]. The reification of race as a natural category can in turn re-entrench systems of racial stratification which give rise to real health and social inequalities between different groups [62, 110]. The unquestioning use of racial categories in scientific research can also lead to misplaced conclusions. For example, the use of race as a natural category can obfuscate the environmental, social, and structural factors that contribute to health disparities [29, 110] and lead to the racialization of certain diseases [19, 30, 37]. The genomic era in particular has given rise to new concerns regarding the use of race in technologies and research. Race-based pharmaceuticals have been critiqued for their part in reifying racial categories as markers of biological difference [59, 63]. Genetic ancestry testing has been criticized for its potential to promote biological essentialilsm and reinforce race privilege amongst those already experiencing it [15,

[3] In a sense, the "observed race/appearance-based" dimension and the phenotype dimension are collapsed for computer vision systems, because a computer vision system can only "see" the objective and phenotypical.

Dimension	Description	Typical Measurement	Outcomes it may be appropriate for studying
Racial identity	Subjective self-identification, not limited by pre-set options.	Open-ended self-identification question	Political mobilization; assimilation; social networks; voting; residential decision-making; attitudes
Racial Self-Classification	The race you check on an official form or survey with constrained options (e.g. the Census)	Closed-ended survey question	Demographic change; vital statistics; disease and illness rates
Observed Race	The race others believe you to be	Interviewer classification	Discrimination; socioeconomic disparties; residential segregation; criminal justice indicators; health care/service provision
- Appearance-Based	Observed race based on readily observable characteristics	Interviewer classification with instructions to classify on first observation	Racial profiling; discrimination in public settings
- Interaction-Based	Observed race based on characteristics revealed through interaction (e.g. language, accent, surname)	Interview classification with instructions to classify after interaction or survey	Workplace discrimination; housing discrimination; language/accent-based discrimination
Reflected Race	The race you believe others assume you to be	A question such as "What race do most people think you are?"	Self-identification processes; perceived discrimination
Phenotype	Racial appearance	Usually interviewer classification, but also "objective" characteristics such as: skin color; hair texture and color; nose shape; lip shape; eye color	Discrimination; socioeconomic disparties; residential segregation; criminal justice indicators; health care/service provision

Table 1: Multiple dimensions of race. Reproduced and amended from [95]. Note that we have excluded "racial ancestry" from this table. Genetics, biomedical researchers, and sociologists of science have criticized the use of "race" to describe genetic ancestry within biomedical research [40, 49, 84, 122], while others have criticized the use of direct-to-consumer genetic testing and its implications for racial and ethnic identification [15, 91, 113].

98]. Finally, bioinformatics tools and research practices themselves contribute to the essentialization of race [10, 42, 43, 51].

There has also been significant debate about the appropriate use of racial variables in studies aimed at identifying causal effects [46, 54, 55, 58, 67]. Objections to the use of race as a causal variable are often couched in terms of manipulability, embodied by the slogan "no causation without manipulation" [54]. These arguments often point to the physical impossibility of manipulating race, thus precluding its use as a causal variable. Several additional, more sociologically-grounded objections to the use of race as a causal variable have emerged within social statistics [55], anti-discrimination legal scholarship [67], and health research [46]. These objections point to the fundamental role race plays in structuring life experiences, making it nonsensical to talk about two individuals being identical, save for race. Stated another way, attempts to isolate the treatment effect of race are often based on a "sociologically incoherent conception of what race references" [67].

Several scholars have pointed to the significant social consequences that arise when race is misconstrued as a variable that can produce causal effects [3, 46, 58, 124]. First, without proper contextualization of results, there is a tendency to misattribute the causal mechanisms of difference the racial categories themselves [3, 46, 124], further reifying race as a natural category. Second, misidentifying race, rather than racial stratification, as the root cause of social and health disparities can lead to misplaced conclusions and ineffective public policy interventions [46].

4.2 Critical race methodologies

These critiques have given rise to a host of critical race methodologies informing the use of racial categories within these respective disciplines. To begin, the choice to use racial categories at all should be carefully examined, particularly in biological contexts. Despite widespread agreement that racial categories do not describe genetically distinct populations [84], genomics researchers have grappled for over a decade with the utility and harm of using racial categories to frame research studies and communicate scientific results. The context of use is key to assessing the appropriate use of racial categories. For example, when adopted as a social or political category, race has utility in genomics, e.g. to document biological differences that result from processes of racial stratification. However, when disconnected from their social and political histories, racial categories have no place in biological research [30, 90, 122]. Genetics, biomedical researchers, and sociologists of science have emphasized the importance of distinguishing ancestry from racial taxonomies within biomedical research [40, 49, 84] and have warned against the dangers of of using racial categories to describe genetic variation [14, 30, 122].

As noted above in our discussion of Benthall and Haynes, race cannot and should not be done away with entirely. The key challenge here is to "denaturalize without dematerializing it" [78] by recognizing race as a multidimensional, relational, and socially situated construct. This begins with centering the process of conceptualizing and operationalizing race, critically assessing the choice of categories and measurement schemes, and fully articulating and justifying these decisions in academic communications [46, 60, 71, 77, 83].

In the context of causal studies, several methodologies have been proposed to mitigate the risks of reifying race as an entity that produces causal effects. For example, some researchers have considered studies that manipulate properties associated with race, such as names [13] [4]. Sen and Wasow's 'bundle of sticks' framework theorizes race as a composite variable that can be disaggregated into constitutive elements, some of which can be manipulated [109].

Epidemiological and public health researchers have increasingly shifted their practices towards the study of social determinants of health, and in particular, the multiple ways racism affects health outcomes [41, 61, 71, 117, 119, 120]. This shift from studying effects of *race* to the effects of *racism* is a central component of critical race methodologies emerging within other disciplines. For instance, Ford and Hawara propose a methodology for the operationalization of race within health equity studies as a multidimensional, context-specific variable with a relational component to explicitly capture the effects of racial stratification on health outcomes [34].

Moreover, appropriately contextualized descriptive studies can be an important tool for identifying and understanding patterns of inequality. When race is utilized as a descriptive category, the choices about what categories to use and how to assign individuals to categories significantly impacts the results of analysis [18, 102, 114, 115]. For example, Howell and Emerson compare the effectiveness of five different operationalizations of race in predicting different measures of social inequality. The study found each operationalization told a different story, pointing to the importance of critically evaluating racial categories and measurement schemes, as well as the importance of articulating these decisions in the communication of results [56].

In recent years, anti-racist efforts have emerged within public health that more thoroughly integrate critical race theory into methodologies and discourse [35, 36]. While earlier studies of the social determinants of health tend to center individual and interpersonal racialized experiences, this new paradigm shifts focus to the institutional and structural conditions that shape racial disparities in health [44, 45, 110, 118]. Anti-racist practice also emphasizes the importance of researchers recognizing their privilege and positionality and the way this might shape their practice. For example, the Public Health Critical Race Praxis offers a self-reflexive and race-conscious research methodology that centers discourse and practice around the perspectives of socially marginalized groups, building on community-based participatory approaches [35].

5 IMPLICATIONS FOR USING RACE IN ALGORITHMIC FAIRNESS RESEARCH

We observe several widespread tendencies when using race within algorithmic fairness research. First, frameworks for describing and mitigating unfairness adopt a simplistic conceptualization of race as a single dimensional variable that can take on a handful of values. This simplification erases the social, economic, and political complexity of the racial categories. Further, methodologies built upon this conceptualization of race tend to treat groups as interchangeable, obscuring the unique oppressions encountered by each group. We argue that this framing limits the effectiveness of fairness analysis and interventions and risks reifying racial categories in the process.

Second, the process of conceptualizing and operationalizing race for the purposes of studying or mitigating different aspects of algorithmic fairness has – with the exceptions noted above – received little attention. Moreover, racial categories are frequently adopted with little attention given to the histories of the categories or suitability of the categories for the fairness assessment.

We discuss the importance of focusing on categories and measurement processes, and fully articulating these decisions in communication of results. Lastly, we discuss the use of disaggregated analysis and the implications for incorporating a more nuanced understanding of measurement for these analyses.

5.1 Race and group fairness

Group fairness criteria represent a class of algorithmic fairness definitions predicated on clearly defined subgroups in the dataset. "Fairness" is obtained by equalizing a particular statistic, or set of statistics, of the classifier across groups. Many different group fairness criteria have been proposed. For example, demographic parity requires equal rates of positive prediction [33] across groups. Equality of odds [50], also referred to as equalized mistreatment [123], requires equal false positive and false negative rates across groups. Equal opportunity [50] restricts this requirement to only a single value of the true outcome. Calibration (or test fairness) [23, 66] requires the actual outcome to be independent of protected attributes, conditioned on estimated outcome.

While limitations of group fairness criteria have been previously explored [26, 47], here we focus our examination on the conceptualization of race embedded within this framework. By abstracting racial categories into a mathematically comparable form, group-based fairness criteria deny the hierarchical nature and the social, economic, and political complexity of the social groups under consideration. Most notably, critical race theorists and Black feminist thinkers have criticized group fairness approaches for their underlying ideal, liberal approaches to ameliorating past harms.

Black feminist cultural geographer Katherine McKittrick, following Patricia Hill Collins, discusses how such approaches have resulted in a 'flattened geography' which obscures the unique oppressions encountered by Black women, instead treating oppressed social groups as interchangeable [25, 79]. Ladson-Billings and Tate, in their critical race critique of liberal education paradigms, explain that the tension between and even among different racial groups are not adequately examined or understood and assume "that all difference is both analogous and equivalent" [68]. Political philosopher

[4] However, Kohler-Hausmann has been critical of such "audit" studies for not being able to isolate the treatment effect for race, although they do, in some circumstances, "provide evidence of a constitutive claim that grounds a thick ethical evaluation" [67, pp. 33-34]

Charles W. Mills notes that the liberal underpinnings of both theoretical and methodological approaches towards equalization have historically tended to have significant detrimental consequences for racial minorities and only serves to advance a white "racial contract" [80].

In short, group fairness approaches try to achieve sameness across groups without regard for the difference between the groups. Group fairness offers an incomplete version of what a race-conscious policy would be. This treats everyone the same from an algorithmic perspective without acknowledging that people are not treated the same. As political philosopher Elizabeth Anderson notes, "[s]tandard conceptions of distributive equality, embodied in ideas of equality of opportunity... fail to consider how such opportunities build in group hierarchy" [4].

5.2 Conceptualizing and operationalizing race

The question of how best to operationalize race for the purposes of studying or mitigating different aspects of algorithmic unfairness has received little attention. By discounting the considerations that go into operationalizing race, the scope and effectiveness of subsequent analysis and interventions fail to interrogate how the particular operationalization and measurement affect a given outcome. For instance, referring to Table 1, while observed race may be more important for studying discrimination, self-identification may be more suitable for studying identity formation and voting behavior.

We suggest centering the process of conceptualizing and operationalizing race when working with racial variables. In particular, we urge algorithmic fairness researchers to critically evaluate existing racial schemas. Because of data limitations, we are most often bound to the categories provided by census categories or other taxonomies which stem from bureaucratic processes. We know from the histories outlined above that these categories are unstable, contingent, and rooted in racial inequality.

Following recent work around measurement and fairness [5, 57], we suggest taking seriously the problems of measurement modeling when considering racial variables. Race, like fairness, is itself a contested concept, and it follows that adopting a multidimensional view is one strategy forward for approaching this. Recently, Roth has called more broadly for a "sociology of racial appraisals" in which we better theorize how observed race has changed over time, and how these changes influence norms of classification [97]. Given that most, if not all, of algorithmic fairness research is ostensibly concerned with anti-discrimination, this charge should be taken seriously.

As a helpful analogue, a growing body of survey research has highlighted the impact that racial category and measurement choices can have on outcomes being measured [102–104]. Similarly, we emphasize that the various choices that go into the operationalization of race for the purposes of fairness-informed analysis or interventions significantly impact the result. Therefore, we need to be transparent about definitions, categories, measurements, and motivations for racial schema used in research and publication. If possible, multiple measures of race should be collected. In addition, measurement of race should be considered as an empirical problem in its own right.

In domains where the categorization is based solely on observable characteristics, e.g. image datasets are annotated by third-parties, care needs to be taken to ensure race is not reduced to phenotype. This consideration can inform how the subgroups themselves are defined (e.g. shifting conceptually from racial categories to categories defined along phenotypic lines) and the way in which the dataset and analysis results are communicated with the larger research community (e.g. detailing procedures for determining categories and category membership, ensuring a distinction is made between phenotypically-determined categories and high-level racial categories). In the context of disaggregated evaluations in computer vision domains, the FAT* community has already seen a marked shift towards defining groups based on phenotypic properties rather than racial categories, thanks to the pioneering work of Buolamwini and colleagues.

However, we also note that shifting to phenotypically-determined categories does not provide a full-stop solution. Given the dark history of physiognomy and phrenology [121], in particular their role in promoting scientific racism and eugenics, researchers should be careful when imposing categories based on, for example, facial landmarks (e.g. IBM's diversity in faces dataset has labelled facial dimensions, proportions, etc.). Doing so risks devolving into using phenotypical features as inputs for a predictive models for individual characteristics and internal psychological states. Furthermore, while phenotype can be used as a means to understand processes of discrimination, practitioners *should not* equate race to phenotype, not to mention other biological markers such as genomes.

5.3 Disaggregated analysis

Frameworks for algorithmic audit studies [22, 92] and standardized model reporting guidelines [81] have highlighted the importance of reporting model statistics disaggregated by groups defined along cultural, demographic, or phenotypic lines. In contrast to group fairness definitions, which are often employed as ideals to be met, disaggregated analysis operates at a descriptive level. These analyses should begin from a pragmatist understanding of differences and interrogate the most salient aspects of race to be considered for particular technologies. For instance, in their audits of facial analysis technologies, Buolamwini and her colleagues concentrate on the phenotypical dimensions of race to understand the disparities of these vision-based systems[5]. Results of disaggregated analysis depend heavily on the choice of categories and racial measurements used.

Within survey research, sociologists of race have investigated the utility of different racial categories for the purposes of understanding markers of social inequality. Howell and Emerson find that a five-fold measure of self-identified race explains more variation in measures of inequality in income, housing, and health in the US [56]. Saperstein and Penner find that racial categorization operates both as an input and an output to racial inequality. Individuals who experience an increase in social position (e.g. increased income) may be more likely to "lighten" themselves in survey responses, while those who become more disadvantaged (e.g. experiencing unemployment or incarceration) may "darken" themselves [89, 105]. The implication here is that it is necessary to understand the way

[5] It should be noted that they do not take a similar approach towards gender.

that a particular racial dimension manifests in measurement and how it relates to the sociotechnical system under consideration. We encourage algorithmic fairness researchers to explore how different racial dimensions and their attendant measurements might reveal different patterns of unfairness in sociotechnical systems.

5.4 Limits of the algorithmic frame

Before undertaking the project of systematizing social groups based on race, we should begin by interrogating why race is a relevant factor in the analysis of the system to begin with. Here, it is critical to expand the scope of analysis beyond the algorithmic frame [108] and interrogate how patterns of racial oppression might be embedded in the data and model and might interact with the resulting system. We urge researchers to question the following: is delineating data along racial lines critical to understanding patterns of fairness produced or reproduced by the system?

In many cases, severe fairness concerns are evident prior to any quantitative race-based analysis of the system's outputs. For example, Rashida Richardson and colleagues recently published an extensive survey of the "dirty data" used to develop predictive policing systems in the US [94]. The report reveals that in nine out of the thirteen jurisdictions reviewed, predictive policing systems ingested data that was generated while the jurisdiction was under investigation for corrupt, racially-biased, or otherwise illegal policing practices. This work represents an analysis centered not around *race* but around *racially disparate policing patterns*. This echoes much of the work from within health and biomedical sciences which emphasizes a racism effect rather than a race effect.

5.5 Centering perspectives of marginalized groups

Lastly, interventions in the vein of algorithmic fairness need to consider the problems of the racially oppressed based on their view of injustices. Political philosopher Elizabeth Anderson proposes that we think about injustices from what she calls a *non-ideal standpoint methodology* [4]. The methodology is non-ideal insofar as that it views injustices from the point of view of the aggrieved groups and the particular harms these groups are facing. The methodology is "standpoint"-based insofar as we should begin from the perspectives of oppressed groups. Here, we can draw lessons from public health scholars who prioritize community-based participatory approaches. These methodologies, which are grounded in critical race theory and feminist standpoint epistemology, take lived experiences of marginalized groups as a valuable and essential source of knowledge.

6 CONCLUSION

In this article, we have pulled back the frame from the algorithms involved in algorithmic fairness themselves and focused on the categories which constitute the "protected classes" used within these frameworks. We focus on race, given the critical nexus of race, technology, and inequality, and the prevalence of race in algorithmic fairness scholarship. We traced the histories of classification, eugenics, and state-sponsored category creation, and warned of the tendency to reify and naturalize racial categories which were the product of long histories of inequality. We then noted the multidimensionality of race and considered lessons from disciplines which have been dealing with this issue for decades. Lastly, we review the algorithmic fairness literature in light of these considerations. We highlight limitations of algorithmic fairness methodologies that adopt a simplistic and de-contextualized understanding of race and outline several suggestions for algorithmic fairness researchers moving forward.

Most in the FAT* community are what Ruha Benjamin calls the "overserved" [11]. In this light, racial classifications need to be interrogated from the perspective of who they serve. Who is doing the categorizing and for what purpose? As in the beginning of this article, we echo J. Khadijah Abdurahman, "it is not just that classification systems are inaccurate or biased, it is who has the power to classify, to determine the repercussions / policies associated thereof and their relation to historical and accumulated injustice." Racial classifications can be an important tool in post-hoc disaggregation analysis. However, we need to know how to disaggregate, how to contextualize categories of disaggregation, and who that categorization serves.

ACKNOWLEDGEMENTS

Thanks to Anna Lauren Hoffmann, Ben Hutchinson, Emma Kaywin, Issa Kohler-Haussman, Joan Fujimura, Lauren Mangels, Simone Wu, and the three anonymous FAT* reviewers for helpful comments.

REFERENCES

[1] J. Khadijah Abdurahman. Fat* be wilin'. 2019.
[2] Syed Mustafa Ali. A brief introduction to decolonial computing. XRDS: Crossroads, The ACM Magazine for Students, 22(4):16–21, 2016.
[3] Walter R Allen, Susan A Suh, Gloria González, and Joshua Yang. Qui bono? explaining–or defending–winners and losers in the competition for educational achievement. In Tukufu Zuberi and Eduardo Bonilla-Silva, editors, White logic, white methods : racism and methodology, chapter 13, pages 217–237. Rowman & Littlefield Publishers, 2008.
[4] Elizabeth Anderson. Toward a non-ideal, relational methodology for political philosophy: Comments on schwartzman's "challenging liberalism". Hypatia, 24(4):130–145, 2009.
[5] McKane Andrus and Thomas K Gilbert. Towards a just theory of measurement: A principled social measurement assurance program for machine learning. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pages 445–451. ACM, 2019.
[6] Julia Angwin and Jeff Larson. Propublica responds to company's critique of machine bias story. ProPublica, 29, 2016.
[7] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. ProPublica, 2016.
[8] Stanley R. Bailey, Mara Loveman, and Jeronimo O. Muniz. Measures of "race" and the analysis of racial inequality in brazil. Social Science Research, 42(1):106 – 119, 2013.
[9] Chelsea Barabas, Karthik Dinakar, Joichi Ito, Madars Virza, and Jonathan Zittrain. Interventions over predictions: Reframing the ethical debate for actuarial risk assessment. arXiv preprint arXiv:1712.08238, 2017.
[10] Jan Baren-Nawrocka. The bioinformatics of genetic origins: how identities become embedded in the tools and practices of bioinformatics. Life Sciences, Society and Policy, 9:7, 09 2013.
[11] Ruha Benjamin. Race After Technology: Abolitionist Tools for the New Jim Code. John Wiley & Sons, 2019.
[12] Sebastian Benthall and Bruce D. Haynes. Racial categories in machine learning. In Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19, 2019.
[13] Marianne Bertrand and Sendhil Mullainathan. Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. American Economic Review, 94(4):991–1013, September 2004.
[14] Catherine Bliss. Racial taxonomy in genomics. Social Science & Medicine, 73(7):1019 – 1027, 2011.

[15] Deborah A Bolnick, Duana Fullwiley, Troy Duster, Richard S Cooper, Joan H Fujimura, Jonathan Kahn, Jay S Kaufman, Jonathan Marks, Ann Morning, Alondra Nelson, et al. The science and business of genetic ancestry testing. Science, 318(5849):399–400, 2007.

[16] Eduardo Bonilla-Silva and Tukufu Zuberi. Toward a definition of white logic and white methods. In Tukufu Zuberi and Eduardo Bonilla-Silva, editors, White logic, white methods: racism and methodology, chapter 1, pages 3–27. Rowman & Littlefield Publishers, 2008.

[17] Geoffrey C. Bowker and Susan Leigh Star. Sorting Things Out. MIT Press, 2000.

[18] Jenifer Bratter and Bridget K Gorman. Does multiracial matter? a study of racial disparities in self-rated health. Demography, 48:127–52, 02 2011.

[19] Lundy Braun, Anne Fausto-Sterling, Duana Fullwiley, Evelynn Hammonds, Alondra Nelson, William Quivers, Susan Reverby, and Alexandra E Shields. Racial categories in medical practice: How useful are they? PLoS medicine, 4:e271, 10 2007.

[20] Andre Brock. Life on the wire: Deconstructing race on the internet. Information, Communication & Society, 12(3):344–363, 2009.

[21] Simone Browne. Dark matters: On the surveillance of blackness. Duke University Press, 2015.

[22] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In FAT*, pages 77–91, 2018.

[23] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5 2:153–163, 2017.

[24] Wendy Hui Kyong Chun. Introduction: Race and/as Technology; or, How to Do Things to Race. Camera Obscura: Feminism, Culture, and Media Studies, 24(1 (70)):7–35, 05 2009.

[25] Patricia Hill Collins. Fighting words: Black women and the search for justice, volume 7. U of Minnesota Press, 1998.

[26] Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. In arXiv:1808.00023, 2018.

[27] Matthew Desmond and Mustafa Emirbayer. What is racial domination? Du Bois Review: Social Science Research on Race, 6(2):335–355, 2009.

[28] William Dieterich, Christina Mendoza, and Tim Brennan. COMPAS risk scales: Demonstrating accuracy equity and predictive parity. Northpoint Inc, 2016.

[29] Denise J. Drevdahl, Debby A. Philips, and Janette Y. Taylor. Uncontested categories: the use of race and ethnicity variables in nursing research. Nursing Inquiry, 13(1):52–63, 2006.

[30] Troy Duster. Race and reification in science. Science, 307(5712):1050–1051, 2005.

[31] Dave Elder-Vass. Towards a realist social constructionism. Sociologia, problemas e práticas, (70):9–24, 2012.

[32] Diana Elliott, Rob Santos, Steven Martin, and Charmaine Runes. Assessing miscounts in the 2020 census. 2019.

[33] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '15, pages 259–268, New York, NY, USA, 2015. ACM.

[34] Chandra Ford and Nina Harawa. A new conceptualization of ethnicity for social epidemiologic and health equity research. Social science & Medicine, 71:251–8, 07 2010.

[35] Chandra L. Ford. Public health critical race praxis: An introduction, an intervention, and three points for consideration. Wisconsin Law Review, 3:477–491, 01 2016.

[36] Chandra L. Ford and Collins Airhihenbuwa. Critical race theory, race equity, and public health: Toward antiracism praxis. American Journal of Public Health, 100 Suppl 1:S30–5, 02 2010.

[37] Morris W. Foster and Richard R Sharp. Race, ethnicity, and genomics: social classifications as proxies of biological heterogeneity. Genome research, 12 6:844–50, 2002.

[38] Michel Foucault. The order of things. Routledge, 2005.

[39] Marion Fourcade and Kieran Healy. Seeing like a market. Socio-Economic Review, 15(1):9–29, 2016.

[40] Joan H Fujimura, Deborah A Bolnick, Ramya Rajagopalan, Jay S Kaufman, Richard C Lewontin, Troy Duster, Pilar Ossorio, and Jonathan Marks. Clines without classes: How to make sense of human variation. Sociological Theory, 32(3):208–227, 2014.

[41] Mindy Fullilove. Comment: Abandoning "race" as a variable in public health research - an idea whose time has come. American Journal of Public Health, 88:1297–8, 10 1998.

[42] Duana Fullwiley. The molecularization of race: Institutionalizing human difference in pharmacogenetics practice. Science as Culture, 16(1):1–30, 2007.

[43] Duana Fullwiley. The biologistical construction of race. Social studies of science, 38:695–735, 11 2008.

[44] Jennifer Garcìa and Mienah Sharif. Black lives matter: A commentary on racism and public health. American Journal of Public Health, 105:e1–e4, 06 2015.

[45] Gilbert Gee and Devon Payne-Sturges. Environmental health disparities: A framework integrating psychosocial and environmental concepts. Environmental health perspectives, 112:1645–53, 01 2005.

[46] Laura E. Gomez and Nancy Lopez, editors. Mapping Race: Critical Approaches to Health Disparities Research. New Brunswick, NJ, Rutgers University Press, 2013.

[47] Ben Green and Lily Hu. The myth in the methodology: Towards a recontextualization of fairness in machine learning. In ICML 2018, 2018.

[48] Ian Hacking. The social construction of what? Harvard university press, 1999.

[49] Nina T Harawa and C. Lawrence Ford. The foundation of modern racial categories and implications for research on black/white disparities in health. Ethnicity & Disease, 19 2:209–17, 2009.

[50] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16, pages 3323–3331, USA, 2016. Curran Associates Inc.

[51] Alana Helberg-Proctor, Anja Krumeich, Agnes Meershoek, and Klasien Horstman. The multiplicity and situationality of enacting 'ethnicity' in dutch health research articles. BioSocieties, 12 2017.

[52] Michael Herzfeld. The social production of indifference. University of Chicago Press, 1993.

[53] Anna Lauren Hoffmann. Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. Information, Communication & Society, 22(7):900–915, 2019.

[54] Paul W. Holland. Statistics and causal inference. Journal of the American Statistical Association, 81(396):945–960, 1986.

[55] Paul W. Holland. Causation and race. In Tukufu Zuberi and Eduardo Bonilla-Silva, editors, White logic, white methods : racism and methodology, chapter 5, pages 93–109. Rowman & Littlefield Publishers, 2008.

[56] Junia Howell and Michael Emerson. So what "should" we use? evaluating the impact of five racial measures on markers of social inequality. Sociology of Race and Ethnicity, 3:14–30, 5 2017.

[57] Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness. 2019.

[58] Angela James. Making sense of race and racial classification. In Tukufu Zuberi and Eduardo Bonilla-Silva, editors, White logic, white methods : racism and methodology, chapter 2, pages 31–45. Rowman & Littlefield Publishers, 2008.

[59] Jonathan Kahn. Misreading race and genomics after bidil. Nature genetics, 37:655–6, 08 2005.

[60] Judith B. Kaplan and Trude Bennett. Use of Race and Ethnicity in Biomedical Publication. JAMA, 289(20):2709–2716, 2003.

[61] Jay Kaufman and Richard Cooper. Commentary: Considerations for use of racial/ethnic classification in etiologic research. American journal of epidemiology, 154:291–8, 09 2001.

[62] Jay S Kaufman. How inconsistencies in racial classification demystify the race construct in public health statistics. Epidemiology, 10 2:101–3, 1999.

[63] Shannon Kelly and Yashwant Pathak. Race and Ethnicity: Understanding Difference in the Genome Era, pages 71–87. 07 2018.

[64] Ibram X Kendi. Stamped from the beginning: The definitive history of racist ideas in America. Random House, 2017.

[65] Akil Kokayi Khalfani, Tukufu Zuberi, Sulaiman Bah, and Pali J Lehohla. Race and population statistics in south africa. In Tukufu Zuberi and Eduardo Bonilla-Silva, editors, White logic, white methods: racism and methodology, chapter 4, pages 63–92. Rowman & Littlefield Publishers, 2008.

[66] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In ITCS, 2017.

[67] Issa Kohler-Hausmann. Eddie Murphy and the Dangers of Counterfactual Causal Thinking About Detecting Racial Discrimination. Northwestern University Law Review, 113(5), 2019.

[68] Gloria Ladson-Billings and William F Tate IV. Toward a critical race theory of education. Teachers College Record, 97:47–68, 1995.

[69] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How we analyzed the compas recidivism algorithm. ProPublica, 9, 2016.

[70] Catherine Lee. "race" and "ethnicity" in biomedical research: How do scientists construct and explain differences in health? Social Science & Medicine, 68(6):1183 – 1190, 2009.

[71] S Lin and J Kelsey. Use of race and ethnicity in epidemiologic research: Concepts, methodological issues, and suggestions for research. Epidemiologic reviews, 22:187–202, 02 2000.

[72] Adam Liptak. Supreme court leaves census question on citizenship in doubt. The New York Times, 2019.

[73] Ian Haney López. White by law: The legal construction of race. NYU Press, 2006.

[74] Mara Loveman, Jeronimo Muniz, and Stanley R. Bailey. Brazil in black and white? race categories, the census, and the study of inequality. Ethnic and Racial Studies, 35, 08 2012.

[75] Neda Maghbouleh. The Limits of Whiteness: Iranian Americans and the Everyday Politics of Race. Stanford University Press, 2017.

[76] Douglas S Massey and Nancy A Denton. American apartheid: Segregation and the making of the underclass. Harvard University Press, 1993.

[77] Vickie Mays, Ninez Ponce, Donna Washington, and Susan Cochran. Classification of race and ethnicity: Implications for public health. Annual Review of Public Health, 24:83–110, 02 2003.

[78] Amade M'charek. Beyond fact or fiction: On the materiality of race in practice. Cultural Anthropology, 28:420–442, 07 2013.

[79] Katherine McKittrick. Demonic Grounds: Black Women and the Cartographies of Struggle. University of Minnesota Press, ned - new edition edition, 2006.

[80] Charles W Mills. The racial contract. Cornell University Press, 1997.

[81] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In FAT, 2019.

[82] Ellis P Monk Jr. The cost of color: Skin color, discrimination, and health among african-americans. American Journal of Sociology, 121(2):396–444, 2015.

[83] Edward Morris. Researching race: Identifying a social construction through qualitative methods and an interactionist perspective. Symbolic Interaction, 30:409–425, 08 2007.

[84] Karama C. Neal. Use and Misuse of 'Race' in Biomedical Research. Online Journal of Health Ethics, 5(1), 2008.

[85] Safiya Umoja Noble. Algorithms of oppression: How search engines reinforce racism. nyu Press, 2018.

[86] Michael Omi and Howard Winant. Racial formation in the United States. Routledge, 2014.

[87] Michael A Omi. The changing meaning of race. America becoming: Racial trends and their consequences, 1:243–263, 2001.

[88] Peggy Pascoe. Miscegenation law, court cases, and ideologies of "race" in twentieth-century america. The Journal of American History, 83(1):44–69, 1996.

[89] Andrew M Penner and Aliya Saperstein. Disentangling the effects of racial self-identification and classification by others: the case of arrest. Demography, 52(3):1017–1024, 2015.

[90] Elizabeth Phillips, Adebola Odunlami, and Vence Bonham. Mixed race: Understanding difference in the genome era. Social forces; a scientific medium of social study and interpretation, 86:795–820, 01 2008.

[91] Ramya Rajagopalan and Joan H Fujimura. Making history via dna, making dna from history. Genetics and the unsettled past: The collision of DNA, race, and history, page 143, 2012.

[92] Inioluwa Deborah Raji and Joy Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In AIES, 2019.

[93] Victor Ray. A theory of racialized organizations. American Sociological Review, 84(1):26–53, 2019.

[94] Rashida Richardson, Jason Schultz, and Kate Crawford. Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. New York University Law Review, 2019.

[95] Wendy D Roth. The multiple dimensions of race. Ethnic and Racial Studies, 39(8):1310–1338, 2016.

[96] Wendy D. Roth. Methodological pitfalls of measuring race: international comparisons and repurposing of statistical categories. Ethnic and Racial Studies, 40(13):2347–2353, 2017.

[97] Wendy D Roth. Unsettled identities amid settled classifications? toward a sociology of racial appraisals. Ethnic and Racial Studies, 41(6):1093–1112, 2018.

[98] Wendy D. Roth and Biorn Ivemark. Genetic options: The impact of genetic ancestry testing on consumers' racial and ethnic identities. American Journal of Sociology, 124(1):150–184, 2018.

[99] Jacob S Rugh and Douglas S Massey. Racial segregation and the american foreclosure crisis. American sociological review, 75(5):629–651, 2010.

[100] Phia Salter and Glenn Adams. Toward a critical race psychology. Social and Personality Psychology Compass, 7, 11 2013.

[101] Gary D. Sandefur, Mary E. Campbell, and Jennifer Eggerling-Boeck. Racial and ethnic disparities in health and mortality among the u.s. elderly population. In Anderson NB, Bulatao RA, and Cohen B, editors, Critical Perspectives on Racial and Ethnic Differences in Health in Late Life, pages 53–94. Washington, DC: The National Academies Press, 2004.

[102] Aliya Saperstein. Double-checking the race box: Examining inconsistency between survey measures of observed and self-reported race. Social Forces, 85(1):57–74, 2006.

[103] Aliya Saperstein. Capturing complexity in the united states: which aspects of race matter and when? Ethnic and Racial Studies, 35(8):1484–1502, 2012.

[104] Aliya Saperstein, Jessica M Kizer, and Andrew M Penner. Making the most of multiple measures: Disentangling the effects of different dimensions of race in survey research. American Behavioral Scientist, 60(4):519–537, 2016.

[105] Aliya Saperstein and Andrew M Penner. Racial fluidity and inequality in the united states. American Journal of Sociology, 118(3):676–727, 2012.

[106] James C Scott. Seeing like a state: How certain schemes to improve the human condition have failed. Yale University Press, 1998.

[107] James C Scott. Against the grain: a deep history of the earliest states. Yale University Press, 2017.

[108] Andrew D. Selbst, danah boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19, pages 59–68, New York, NY, USA, 2019. ACM.

[109] Maya Sen and Omar Wasow. Race as a bundle of sticks: Designs that estimate effects of seemingly immutable characteristics. Annual Review of Political Science, 19:499–522, 05 2016.

[110] Abigail A. Sewell. The racism-race reification process: A mesolevel political economic framework for understanding racial health disparities. Sociology of Race and Ethnicity, 2(4):402–432, 2016.

[111] Andrew Smart, Richard Tutton, Paul Martin, George T.H. Ellison, and Richard Ashcroft. The standardization of race and ethnicity in biomedical science editorials and uk biobanks. Social Studies of Science, 38(3):407–423, 2008.

[112] C Matthew Snipp. Racial measurement in the american census: Past practices and implications for the future. Annual Review of Sociology, 29(1):563–588, 2003.

[113] Kim TallBear. Native American DNA: Tribal belonging and the false promise of genetic science. U of Minnesota Press, 2013.

[114] Edward Telles, René D. Flores, and Fernando Urrea-Giraldo. Pigmentocracies: Educational inequality, skin color and census ethnoracial identification in eight latin american countries. Research in Social Stratification and Mobility, 40:39 – 58, 2015.

[115] Edward E. Telles and Nelson Lim. Does it matter who answers the race question? racial classification and income inequality in brazil. Demography, 35(4):465–474, 1998.

[116] Charles Tilly. Coercion, Capital, and European States, AD 990-1992. B. Blackwell, 1990.

[117] David Williams and Pamela Braboy Jackson. Social sources of racial disparities in health. Health affairs (Project Hope), 24:325–34, 03 2005.

[118] David Williams and Chiquita Collins. Racial residential segregation: A fundamental cause of racial disparities in health. Public Health Reports, 116:404–416, 09 2001.

[119] David R. Williams. The concept of race in health services research: 1966 to 1990. Health Services Research, 29(3):261–274, 1994.

[120] David R. Williams and Michelle Sternthal. Understanding racial-ethnic disparities in health: sociological contributions. Journal of Health and Social Behavior, 51 Suppl(Suppl):S15–S27, 2010.

[121] Blaise Agüera y Arcas, Margaret Mitchell, and Alexander Todorov. Physiognomy's new clothes. Medium, 2017.

[122] Michael Yudell, Dorothy Roberts, Rob DeSalle, and Sarah Tishkoff. Taking race out of human genetics. Science, 351(6273):564–565, 2016.

[123] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In Proceedings of the 26th International Conference on World Wide Web, WWW '17, pages 1171–1180, Republic and Canton of Geneva, Switzerland, 2017. International World Wide Web Conferences Steering Committee.

[124] Tukufu Zuberi. Deracializing social statistics: Problems in the quantification of race. The Annals of the American Academy of Political and Social Science, 568:172–185, 2000.

[125] Tukufu Zuberi. Thicker Than Blood: How Racial Statistics Lie. U of Minnesota Press, 2001.