# Racial/Ethnic Categories in AI and Algorithmic Fairness: Why They Matter and What They Represent

Jennifer Mickel
jamickel@utexas.edu
University of Texas at Austin
United States of America

## ABSTRACT

Racial diversity has become increasingly discussed within the AI and algorithmic fairness literature, yet little attention is focused on justifying the choices of racial categories and understanding how people are racialized into these chosen racial categories. Even less attention is given to how racial categories shift and how the racialization process changes depending on the context of a dataset or model. An unclear understanding of *who* comprises the racial categories chosen and *how* people are racialized into these categories can lead to varying interpretations of these categories. These varying interpretations can lead to harm when the understanding of racial categories and the racialization process is misaligned from the actual racialization process and racial categories used. Harm can also arise if the racialization process and racial categories used are irrelevant or do not exist in the context they are applied.

In this paper, we make two contributions. First, we demonstrate how racial categories with unclear assumptions and little justification can lead to varying datasets that poorly represent groups obfuscated or unrepresented by the given racial categories and models that perform poorly on these groups. Second, we develop a framework, CIRCSheets, for documenting the choices and assumptions in choosing racial categories and the process of racialization into these categories to facilitate transparency in understanding the processes and assumptions made by dataset or model developers when selecting or using these racial categories.

## KEYWORDS

racial categories, racialization, algorithmic fairness, race and ethnicity

**ACM Reference Format:**
Jennifer Mickel. 2024. Racial/Ethnic Categories in AI and Algorithmic Fairness: Why They Matter and What They Represent. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3630106.3659050

## 1 INTRODUCTION

The utilization of racial and ethnic categories in the development of datasets and models facilitates the inclusion and documentation of diverse perspectives. Racial and ethnic categories are especially crucial for datasets and models in which race and ethnicity serve as relevant factors, may act as confounding variables, or enable the ability to audit for fairness using race and ethnicity for fairness purposes. For example, understanding the racial and/or ethnic target of hate speech is crucial for understanding the impact of hate speech, as hate speech can differ based on the race and/or ethnicity of the target [55]. Similarly, in health, race is correlated with health outcomes [7], and knowledge of a patient's race and ethnicity can help contextualize the patient's experience and health history [60]. In algorithmic fairness settings, knowledge of an individual's race and ethnicity allows for auditing of existing datasets and systems, and many fairness toolkits, such as Fairlearn, rely on this data [12, 42]. Despite the benefit of race and ethnicity, little justification is provided for the racial and ethnic categories chosen and why these categories are most relevant for a dataset or model's particular domain. Furthermore, even if the choice of racial and ethnic categories is justified, even less discussion of how these racial and ethnic categories are assigned to individuals and what factors influence the racialization of people into these categories is given. Discussion of how people are assigned or racialized into these categories is crucial as the racialization of people into particular racial groups varies based on cultural context [24]. Discussing this racialization process allows for understanding how the cultural context(s) and domain(s) affect people's placement and racialization into racial categories.

The racial and ethnic categorization schema used in datasets and models varies based on numerous factors. Some racial schemas used are binary, as in Black/non-Black, Black/White, and White/non-White, while others use multiple racial categories, as in Asian, Black, Hispanic, and White [1]. The racial and ethnic categories selected determine what racial and ethnic experiences are valued and will be traceable. In the binary setting, this often leads to the exclusion of people not racialized into these groups, and people with multiple racial identities are obscured. In the case of White/non-White, the experiences of non-White individuals are treated similarly since they are in the same category, even though it is evident that the experiences of non-White individuals vary drastically. For example, the experiences of Asians and Blacks within the US cultural context vary immensely [16].

In this paper, we discuss in greater depth the effect of racial categorization choices on datasets and models, and we demonstrate the importance of documenting choices and motivations for racial categories by showcasing how ill-defined racial categories can affect datasets and model performance. Our work is grounded in critical race theory and race and ethnic studies. We apply these disciplines' findings and research to the development of datasets and models



*FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil*

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3659050

utilizing racial and ethnic categories. Previous scholarship on racial categories in algorithmic and AI fairness [1, 10, 29] motivates our work as does existing scholarship discussing documentation frameworks for datasets and models [9, 17, 20, 23, 28, 32, 34, 50, 64]. We extend this work by focusing on how the choice of racial categorization and the racialization of people into the chosen racial categories affects how well-represented people are and, subsequently, dataset quality and model performance. To combat these effects, we develop CIRCSheets, a novel framework, grounded in critical race theory, allowing developers of datasets or models to document their motivations behind why they selected certain racial categories and consider the effects of their choice in racial categories.

## 2 BACKGROUND AND RELATED WORK

### 2.1 Racial Categories: The Status Quo

Racial categories used in datasets and models tend to align with the US cultural context [1, 10, 29]. Abdu et al. [1] identify two main choices for racial categorization: binary and more than two races. When binary racial categorization is chosen, it often operates under a Black/White axis. If the racial classification selected is more than two races, the racial categorizations tend to echo the US census [1, 10]. The common categories used with multiple racial categories were Asian, Black, Hispanic, and White [1].

The use of racial categories in datasets and models can help ensure a wide variety of perspectives are represented and considered. Furthermore, the presence of racial categories aids in analyzing, testing, and auditing datasets and models for disparities between racial groups. Without racial categories, these analyses along the axis of race would be challenging to conduct [69, 76]. Unfortunately, poorly defined racial categories can hinder actualizing these benefits [56, 65]. This can occur if a racial category comprises multiple groups whose experiences of racialization vary because the racial group no longer serves as a meaningful proxy for people's lived experiences within those groups.

An example of a racial category that comprises multiple groups who are racialized differently in the US cultural context is White. Individuals of Middle Eastern and North African (MENA) descent are categorized as White within the US despite many members of MENA not perceiving themselves to be White [47]. Furthermore, within the US cultural context, their lived experience and racialization differ from people of European ancestry [46]. Having MENA as part of the White racial category obfuscates the experiences of members of MENA within datasets and models, preventing researchers from observing disparate health outcomes of this group [5]. Practitioners and researchers cannot see if a model performs poorly on MENA or if a dataset accounts for the experiences of people who are part of MENA. Most existing fairness toolkits require demographic information to audit algorithms, so practitioners who use these tools cannot audit their models for information on how the model performs on MENA [42].

A racial category can obfuscate people within that category when a multiracial ethnicity is treated as a racial category. For example, Latinx is a multiracial ethnicity, and the experiences of Latinxs can vary drastically based on the cultural context they are in and their race. For example, in the US, the experiences of lighter-skinned and darker-skinned Latinxs differ [75]. Darker-skinned Latinxs racialized as Black in the US cultural context experience anti-Black discrimination from Latinxs and Whites [30]. Placing all Latinxs into the Latinx category would obfuscate the experiences of darker-skinned Latinxs and prevent researchers and practitioners from observing whether datasets include darker-skinned Latinxs and if models perform poorly on darker-skinned Latinxs.

### 2.2 Race and Ethnicity

Race and ethnicity, although similar, are two different concepts. Racial groups are differentiated by physical differences in certain social constructs [8, 62]. Whereas ethnic groups, are differentiated based on social practices such as "language, religion, rituals, and other patterns of behavior" [8, 62, 80, pp. 106]. Often, ethnic categories are treated as racial categories, which can pose a problem when an ethnicity is not synonymous with a race, as in the case of panethnicities (defined in Section 3.3). For example, some Afro-Latinx individuals identify or are racialized as Latinx ethnically and Black racially [33]. This can lead to obfuscation for Afro-Latinxs and members of other multiracial panethnicities because it is unclear whether an individual's racial identity takes precedence over their ethnic identity or vice versa.

Race and ethnicity, although they have no biological determinant, have real impacts on people's lives, ranging from their health to education to work [11, 15, 49, 53, 77]. Documenting race and ethnicity within datasets and models allows us to see how models perform on various races and ethnicities and helps audit the model for disparate impact. Furthermore, practitioners can train models using loss functions or other techniques that utilize race to help mitigate the oppression people of various racial and ethnic groups experience. Loss functions, used to train models, can be designed to help fulfill these goals [39]. Without knowledge of race and ethnicity, it is incredibly challenging to audit for disparate performance along the axes of race and ethnicity.

### 2.3 Racialization

Racialization refers to the process by which racial meaning is given to people [54]. Factors of physical difference, such as skin color and eye shape, among others, affect how people are racialized, as do accents [18, 62, 71]. The process of racialization varies depending on cultural context, and relevant features in one context may be irrelevant in another [73]. For example, the racial identification of Latinx adolescents and young adults shifts from adolescence to young adulthood and varies depending on generational time in the US, demonstrating that the process of racialization within the US and Latin American countries varies substantially enough for their racial identities to change [36]. Furthermore, as time spent in the US increases, an individual's racial identity is less likely to shift [36].

Self-racial identification and external racialization differ. For example, the responses of Puerto Ricans and Dominicans to the race question on the 2010 US Census differ drastically, with respondents interpreting the question of race differently and using different aspects of race to answer the question [68]. This leads to racial self-identification that differs from how Puerto Ricans and Dominicans would be racialized based on their phenotype within the US [68]. This is due, in part, to different cultural contexts between the US,

Puerto Rico, and the Dominican Republic [68]. For example, in the Dominican Republic, Black is used to describe Haitians [37]. This leads to the racial self-identification of Hispanics on the US Census racialized as Black in the US cultural context to be a poor proxy for their physical features [72].

Salient features of racialization can differ based on the cultural context one is in. In the US, skin color plays a large role in racializing people into racial categories [51, 62]. In Latin America, physical features other than skin color, such as hair texture and facial structure, play a part in racializing someone as Black, causing Latinx individuals with similar skin tones to be racialized differently due to other physical features such as hair texture and facial features [30]. Utilizing racial categories without discussing how people are racialized prevents us from understanding who comprises these racial categories and what factors affect whether people are racialized into particular categories and can lead to harm if we transpose different understandings of racial categories and racialization.

## 2.4 Racial Categories: Contextual Relevance and History

The choice of racial categories in datasets and models is influenced by an array of sociotechnical factors, ranging from technical factors, such as model limitations, to contextual relevance, such as cultural context [1]. Datasets and models developed within the US cultural context tend to utilize racial groups relevant to the US cultural context but provide little justification for these choices [1]. Sometimes the US census is used as justification, as in Andrus et al. [3] or prior work, as in Yang et al. [79], but most position cultural context as a sufficient justification of racial categories, as in Borradaile et al. [14].

Race has been central to political life in the United States [62]. This is evident through political discourse, legal history, and the US Census [62]. The census has been used as a tool to encode these values [1], which is evident when observing the history of racial categories within the US Census. As an example, the Census of 1890 had four categories to classify people with African ancestry out of a total of eight categories [70]. This preoccupation with blackness in 1890 reflects the political climate within the southern states at the time [43]. The US Census of 1960 also reflects the political climate of the time, as Hawaii became a state in 1959 and Hawaiian and part-Hawaiian were added as racial categories to the US Census for the first time [43, 61]. Observing the racial categories in the census over the years showcases how race within the US cultural context has shifted. Before 1860, the racial categories the census included were along the axis of Black and White, but as Asian immigrants immigrated to the US, Asian racial categories were added [31].

Racialization for certain groups varies depending on the context and domain. For example, the racialization of Filipinos varies by context [58]. Some Filipinos identify culturally as Latinx rather than Asian, but within educational contexts, they tend to be treated as Asian rather than Latinx [57, 58]. This is seen in the literature for some studies racialize Filipinos as Asian, as in Baluran et al. [6] and Irizarry et al. [35] while others racialize Filipinos as Hispanic, as in Treviño [74].

In addition to the context associated with the domain one operates in, racialization is affected based on the cultural context [24]. For example, the experience of Central-East European immigrants differs between the UK and Japan [24]. In addition to this, the experiences of certain groups within a racial category vary. For example, East Asians and South Asians are both racialized as Asian, yet their experiences differ, which leads Americans of Chinese descent to have a higher life expectancy than Americans of South Indian descent [6].

Racial categories also differ based on country. Farquharson [25] discusses the racial formation of racial categories in the US, South Africa, and Australia, all of which are settler colonial states and identify race along a Black/White axis. Despite this, within each cultural context, people are racialized into the Black category differently. In South Africa, people of African ancestry who are mixed are racialized as colored, while in the US, they would be considered Black [22, 41]. In Australia, the Aboriginal peoples are racialized as Black, while in the US, they would not be [25]. Lack of justification regarding racial categories prevents critical analysis of the sociological foundation of racial categories.

Differing notions of racialization have led to variations between computer vision datasets using similar racial categories in similar cultural contexts that lead each dataset to have differing racial systems [40]. Khan and Fu [40] have identified variations in racialization within computer vision datasets, leading to inconsistent racial systems across datasets developed in similar cultural contexts.

## 2.5 Researcher Justifications

Abdu et al. [1] identify five existing categories of racial category justification in the algorithmic fairness literature. Researcher justifications fall under data availability, technical factors, appeals to prior scientific work, epistemic concerns, and contextual relevance [1]. The first two categories of justifications, data availability and technical factors, focus on limiting factors that affect racial category justification. Data availability affects the racial categories researchers can choose because the choice of racial categories was made earlier during the data curation process. Furthermore, researchers and practitioners must rely on the information regarding racial categories and racialization provided with the data. In many cases, this means no information is provided [1]. Technical factors can affect the racial categories chosen because the model or algorithm may require or be limited to a certain number of features, as in the case of Friedler et al. [27] where their model required a binary racial category as the algorithm's sensitive attribute.

The last three categories of justification appeal to prior scientific work, epistemic concerns, and contextual relevance, focus on justifications related to the goal of the dataset or model and the domain(s) and cultural context(s) in which the dataset and model will be used. Appeals to prior scientific work utilize existing literature as justification for the racial categories used [1]. Justifications regarding epistemic concerns centering racial categories with greater scientific rigor, such as describing what features constitute a person's placement into a particular racial category [1]. Cultural context refers to the racial categories that are relevant in particular societies [1]. Oftentimes, there is an assumption of collective understanding that the racial categories chosen are salient for a certain cultural

context. For example, datasets developed in the US cultural context, as in Borradaile et al. [14], will justify their choice of racial categories by saying they are relevant to the US context.

## 3 HOW RACIAL/ETHNIC CATEGORIES CAN AFFECT DATASETS AND MODELS

With the usage of racial and ethnic categories during dataset and model development, it is often unclear who fits into these categories due to the lack of discussion regarding assumptions about who is racialized into these categories. The cultural relevance and demographic makeup of these categories, as well as the multidimensionality of race and ethnicity, can impact a dataset's quality and a model's performance. Section 3.2 demonstrates how different demographic distributions, possible in broad or ill-specified racial and ethnic categories, can affect model accuracy on a group level.

### 3.1 The Effect of Cultural (Ir)relevance

Within race and ethnic studies, it is well-documented that cultural relevance of racial and ethnic categories is crucial and these categories can shift depending on cultural context and time period [13, 19, 21, 44, 48, 62, 63, 70]. Bonilla-Silva [13] discusses how the structure of racial order is shifting in the US from a bi-racial order to a tri-racial order. Pirtle [63] discusses how the state of South Africa created a tri-racial hierarchy, and how this hierarchy affected racialization in South Africa. Thus, cultural relevance is crucial when selecting racial categories as racial categories vary depending on cultural context [25]. If the racial categories selected for a cultural context are irrelevant to the domain(s) and context(s) they will be deployed in, the benefit of racial categories is lost, as racial categories lose their meaning when irrelevant. Khan and Fu [40] demonstrate this as they find that computer vision datasets annotated using similar racial categories lead to varying racial systems. This is, in part, due to the lack of standardization regarding racialization, leading to differing racial systems. These differing racial systems across computer vision datasets lead to challenges when evaluating for fairness criteria [40].

Some racial categories, such as Black, may exist in multiple cultural contexts, but the people placed into this category change depending on the context. A poorly defined definition of Black, which occurs when there is little to no discussion of how people are racialized into the category of Black, can lead to the usage of varying definitions of Black, especially if a dataset or model is used in a variety of cultural contexts. This has occurred in the US where people have been categorized as Black even though they would be racialized as white [70].

To illustrate this effect, imagine a dataset or model is developed for the cultural contexts of the US, South Africa, and Australia, where Black is a culturally relevant racial category [25]. The developers are aware that Black as a racial category exists in each of these contexts and select the racialization process for the Black racial category to be culturally relevant to Australia, which refers to the Aboriginal people as Black [25]. The developers make this selection without conveying the racialization process of people into the Black category. Another group decides to use the dataset or model in the US or South Africa without understanding that people racialized into the Black category within this dataset or model are Aboriginal. This can lead to downstream issues or harm as the Black category is not relevant to the US or South African context since the racialization process differs from that of Australia. To prevent this from occurring, it is crucial to understand how people are racialized into each racial category of a dataset or model to understand if those racialization processes are culturally relevant to the domain(s) users of the dataset or model want to utilize it for.

### 3.2 The Effect of Distribution Shift in Broad Categories

Abdu et al. [1] identify two main choices for racial categorization: binary and more than two races. Previous work using binary racial categorization utilizes Black/White, Black/non-Black, or white/non-White [1]. Non-Black and non-White are broad categories, and the possible sociodemographic distributions can vary drastically. With these racial categorization schemas, it becomes unclear which groups comprise non-Black and non-White. The non-White category could be comprised solely of Latinx individuals or of both Black and Latinx individuals. Understanding the composition of broad racial categories and *who* is included in these categories is crucial. Otherwise, dataset quality and model performance metrics might differ if the distributions within these broad categories shift.

To demonstrate the impact of this, we use the dataset associated with COMPAS, an algorithm used to predict the recidivism risk of defendants, to train a logistic regression classifier using varying distributions of data based on the racial and ethnic categories in the dataset [4][1]. Our logistic regression classifiers are trained race-blind and use a threshold of 0.5. We test the logistic regression classifiers on each demographic group individually, all demographic groups, and the demographic groups trained on. Our results are showcased in Table 1, which demonstrates that performance metrics vary based on the data each logistic regression model was trained on. The overall accuracy for all groups between the classifiers is within 1%, but, per group, the difference between accuracies can range almost three times that for Hispanic and Other and two times that for African American and Caucasian. This means that the choice of racial categorization schema, racial categories, and who is racialized into these categories can have a real effect on whether someone is more likely to be correctly predicted to rescind. The true positive rate varies within 5%, and the false positive rate varies within 7% across all groups. These figures only increase when looking at each group individually. African American, Hispanic, and Other have higher false positive rates to begin with, so individuals in these groups would be more affected by this variation in false positive rates. The positive predictive value and false discovery rate vary by 2.6% for all groups and up to almost double that for Hispanic (4.6%) and Other (5.5%).

This variation also occurs within racial categorization schemas. For White/non-White, the performance metrics can vary around 5% when comparing Everyone, White/Black, White/non-White (Hispanic + Other), and White/non-White (Hispanic), which would all be valid distributions under the White/non-White categorization. Similar variation occurs for Black/non-Black when comparing Everyone, White/Black, Black/non-Black (Hispanic + Other), and Black/non-Black (Hispanic), which would all be valid distributions

[1]Code available here: https://github.com/jennm/racial-categorizations-ai-fairness

| Classifier | Metric | Asian | African American | Caucasian | Hispanic | Native American | Other | All | Groups Trained On |
|---|---|---|---|---|---|---|---|---|---|
| **Everyone** | TPR (%) | 100.0 | 73.6 | 50.6 | 42.2 | 100.0 | 42.9 | 62.9 | 62.9 |
| | FPR (%) | 0.0 | 39.7 | 15.2 | **22.8** | 0.0 | 24.4 | 27.8 | 27.8 |
| | PPV (%) | 100.0 | 66.8 | 70.9 | *59.4* | 100.0 | *54.5* | 67.0 | 67.0 |
| | FDR (%) | 0.0 | 33.2 | 29.1 | **40.6** | 0.0 | **45.5** | 33.0 | 33.0 |
| | Acc (%) | 100.0 | 67.2 | 70.3 | *61.8* | 100.0 | *62.3* | 67.8 | 67.8 |
| **Black/ White** | TPR (%) | 100.0 | **74.2** | **51.7** | 44.4 | 100.0 | 46.4 | **63.9** | 66.3 |
| | FPR (%) | 0.0 | 41.0 | 15.6 | **22.8** | 0.0 | 24.4 | 28.5 | 29.7 |
| | PPV (%) | 100.0 | 66.2 | 70.8 | 60.6 | 100.0 | 56.5 | 66.8 | 67.4 |
| | FDR (%) | 0.0 | 33.8 | 29.2 | 39.4 | 0.0 | 43.5 | 33.2 | 32.6 |
| | Acc (%) | 100.0 | 66.9 | **70.5** | 62.7 | 100.0 | 63.8 | 67.9 | **68.4** |
| **White/ non-White (Hispanic + Other)** | TPR (%) | 100.0 | 71.8 | *44.4* | *35.6* | 100.0 | 42.9 | 59.5 | *42.6* |
| | FPR (%) | 0.0 | 35.1 | 14.4 | *15.8* | 0.0 | *19.5* | 24.4 | 15.2 |
| | PPV (%) | 100.0 | **68.9** | 69.3 | **64.0** | 100.0 | **60.0** | 68.6 | 67.3 |
| | FDR (%) | 0.0 | *31.1* | 30.7 | *36.0* | 0.0 | *40.0* | 31.4 | 32.7 |
| | Acc (%) | 100.0 | **68.5** | *68.2* | 62.7 | 100.0 | 65.2 | **68.0** | 66.9 |
| **White/ non-White (Hispanic)** | TPR (%) | 100.0 | *69.7* | *44.4* | *35.6* | 100.0 | *35.7* | *57.9* | *42.6* |
| | FPR (%) | 0.0 | *34.8* | *12.3* | *15.8* | 0.0 | *19.5* | *23.5* | *13.0* |
| | PPV (%) | 100.0 | 68.5 | **72.5** | **64.0** | 100.0 | 55.6 | **68.9** | **70.9** |
| | FDR (%) | 0.0 | 31.5 | *27.5* | *36.0* | 0.0 | 44.4 | *31.1* | *29.1* |
| | Acc (%) | 100.0 | 67.6 | 69.4 | 62.7 | 100.0 | *62.3* | 67.7 | 68.1 |
| **Black/ non-Black (Hispanic + Other)** | TPR (%) | 100.0 | 73.6 | 49.4 | **46.7** | 100.0 | **50.0** | 63.2 | 69.0 |
| | FPR (%) | 0.0 | 41.3 | 16.5 | 21.1 | 0.0 | 24.4 | 28.8 | 36.7 |
| | PPV (%) | 100.0 | 65.9 | 68.8 | 63.6 | 100.0 | 58.3 | 66.3 | *65.3* |
| | FDR (%) | 0.0 | 34.1 | 31.2 | 36.4 | 0.0 | 41.7 | 33.7 | **34.7** |
| | Acc (%) | 100.0 | 66.5 | 69.1 | **64.7** | 100.0 | 65.2 | *67.4* | *66.1* |
| **Black/ non-Black (Hispanic)** | TPR (%) | 100.0 | 73.9 | 50.0 | **46.7** | 100.0 | **50.0** | 63.6 | **70.7** |
| | FPR (%) | 0.0 | **41.6** | **16.9** | 21.1 | 0.0 | 24.4 | **29.1** | **38.4** |
| | PPV (%) | 100.0 | *65.8* | *68.5* | 63.6 | 100.0 | 58.3 | *66.2* | 65.6 |
| | FDR (%) | 0.0 | **34.2** | **31.5** | 36.4 | 0.0 | 41.7 | **33.8** | 34.4 |
| | Acc (%) | 100.0 | 66.5 | 69.1 | **64.7** | 100.0 | 65.2 | *67.4* | 66.2 |

**Table 1: Recidivism prediction performance is measured by the true positive rate (TPR), false positive rate (FPR), positive predictive value (PPV), false discovery rate (FDR), which is (1 - PPV), and accuracy (Acc). The groups in parentheses next to the category in "Classifiers" refer to the groups the logistic regression models are trained on. For example, Black/non-Black (Hispanic + Other) means that the non-Black category consisted of those in the categories of Hispanic or Other, and the model was trained on Black, Hispanic, and Other data points. The bolded numbers correspond to the classifier with the highest percentage for that particular metric, and the *italicized* numbers correspond to the classifier with the lowest percentage for that particular metric. If something occurs three or more times, it is not bolded or italicized, even if it meets the criteria. Asian and Native American are the same for each classifier, so none of those metrics are bolded or italicized.**

under the Black/non-Black categorization. Even in more specific racial and ethnic categories like Asian, Black, Latinx, Indigenous, Pacific Islander, and White this can transpire, for different distributions of various ethnic groups or racial groups can occur in these categories, which can also lead to variation in dataset quality and model performance.[2]

## 3.3 The Effect of Racial Multi-Dimensionality and Panethnicity

Abdu et al. [1] and Benthall and Haynes [10] discuss the importance of racial categories, and we expand upon this to emphasize the importance of considering both multiracial and panethnic identities. Existing usage of racial categories in datasets and models rarely allows for multiracial and panethnic identities. Due to technical limitations [1], each person is assigned a singular racial category and is rarely assigned more than one racial category. This leads multiracial individuals and their experiences to be obfuscated in either a racial category that comprises part of their racial experience or an 'Other' category where other multiracial individuals are placed,

[2]Further analysis is available in the Appendix.

often with differing experiences of race [26, 66]. This manifests in models as Wolfe et al. [78] demonstrate that multiracial people are more likely to be assigned a racial or ethnic label of a minority group rather than a majority group.

Panethnic identities are, similarly, seldom adequately represented in the racial and ethnic categories used in datasets and models [36]. Panethnicity refers to the identity that forms when different ethnic or tribal groups build institutions and identities across these ethnic groups' boundaries, leading to panethnicities comprised of people of various racial identities [59]. There are numerous panethnicities, and Latinx is an example of a panethnicity [52].

When panethnicities are included as a category in the chosen racial/ethnic categories selected or used by practitioners, the panethnic categories tend to be treated as a racial category regardless of the other racial identities members of panethnic groups may have. This leads the racial identities of members of this panethnicity to be unaccounted for and causes members of a panethnicity to be treated similarly due to their categorization, obfuscating the varying experiences of people that can be associated, in part, with their racial identity [45]. This is readily seen within the US cultural context when Latinx as a category is used to solely represent the experiences of Latinx individuals, negatively affecting Afro-Latinxs, as their identities are obfuscated since often they are unable to select a racial category that best describes their racial identity and experience. Many Afro-Latinxs are not accepted as Latinx by their lighter-skinned peers, leading some Afro-Latinx individuals to find solidarity in Black communities where they feel more accepted [33]. Placing Afro-Latinxs solely in the Latinx category would prevent datasets and models from being able to account for these experiences of Afro-Latinxs.

# 4 CIRCSHEETS: A DOCUMENTATION FRAMEWORK FOR CONSIDERATIONS IN RACIAL CATEGORIZATION SELECTION

We present CIRCSheets, a framework to articulate the choices of racial categories to better position and understand the effect of racial categorization choices made in developing a dataset or model. Previous documentation frameworks do not address racialization processes [9, 17, 20, 23, 28, 32, 34, 50, 64], and CIRCSheets addresses this gap. This framework addresses the concerns outlined in Section 3 by providing questions practitioners should address and considerations they should consider when answering these questions. CIRCSheets was validated through conversations with academics in the fields of Black Studies, Latinx Studies, and Sociology and evaluated to ensure a user would have a sufficient understanding of why certain racial and ethnic categories were selected, who is racialized into each category, and how the curators' lived experiences informed their choices in categories. Thus, CIRCShets allows for an improved understanding of the assumptions and choices made by the users and developers of datasets and models, helping future dataset and model users understand whether the racial categories are relevant to their use case while decreasing the likelihood of misaligned interpretations of the racial categorizations and the racialization processes from the creators of the dataset or model.

## 4.1 Categories

*Considerations*

- Consider how data availability and technical implementation affect how race and ethnicity can be represented in the dataset and/or model.
- Consider the domain(s) for which the dataset or model is developed for and how this affects the racial categories salient to these domain(s) and the racialization process(es).
- Consider how well the chosen racial categories represent the population(s) represented by the dataset or the population(s) affected by the model.

*Documentation Questions*

(1) What are the racial categories utilized?
(2) What is the motivation behind using these racial categories?
(3) Are multiracial ethnic categories utilized?
(4) If multiracial ethnic categories are used, what is the motivation behind using these categories, and are they being treated as racial categories?
(5) Are people who select multiple racial categories considered multiracial? Are people who select one or more ethnic categories and one racial category considered multiracial?
(6) If so, what category are they placed into, and are other people who select multiple different racial and/or ethnic categories also placed into that same category? If not, what category are they placed in, and does ethnicity take priority over race?
(7) For models, what is the technical implementation of the racial and/or ethnic categories?
(8) How do ethnic groups fit into these racial categories?
(9) Can people be obfuscated by these racial categories? If so, do these groups experience erasure and is the model or dataset likely to interact with them?

## 4.2 Racialization

*Considerations*

- Consider what contexts the dataset or model will be used in and how this affects racialization.
- Consider what factors will be used in the racialization process and who determines an individual's racial identity.
- Consider what the most relevant factors of racialization are within the context(s) the dataset or model operates within.

*Documentation questions*

(1) Who determines an individual's racial categorization? Is it the individual?
(2) Are physical characteristics asked of an individual?
(3) Is cultural background asked of an individual?
(4) In what ways could the existing racial information be partial or incorrect? What impact could this have on the dataset or model?
(5) If using an existing dataset and no racialization information exists, what was the source of the dataset, what cultural context was it developed in, and is there any existing scholarship on the racialization choices of that dataset?

## 4.3 Cultural Context

*Considerations*

- Consider how racial identification can change in the chosen cultural context(s) of your dataset or model.
- Within the cultural context(s) the dataset or model operates in, consider what groups experience marginalization and how the choice of racial categories can affect what groups have visibility in the dataset or model.
- For data collection and dataset development, consider what viewpoints associated with racial identification you want to be represented within your dataset.

*Documentation questions*

(1) What cultural context(s) is this dataset or model developed for?
(2) Will this dataset or model be used in different cultural context(s)?
(3) If the dataset or model is used in different cultural context(s) or domains, is there any misrepresentation that can occur due to changes in racialization or racial categories within these different cultural contexts and domains?

## 4.4 Multi-racial and pan-ethnicity

*Considerations*

- Consider how multiracial individuals and multiracial panethnicities are represented within racial categories and whether the representation of these ethnicities can lead to obfuscation between people of different races within those panethnicities.
- Consider representing racial categories and ethnicities separately.
- Consider the representation of multiracial individuals within the dataset or model and whether this reflects their lived experiences within society.
- Consider whether technical limitations influence whether multiracial individuals can be adequately represented within models.

*Documentation questions*

(1) How are multiracial individuals and multiracial panethnicities categorized within the dataset or model?
(2) Can more than one racial or ethnic category be selected?
(3) Do the categories given to panethnic individuals effectively communicate their racial and ethnic identities?
(4) Are there any individuals, such as Afro-Latinxs, who may be *in*adequately represented by the racial categorizations chosen?

## 4.5 Knowledge and Positionality

*Considerations*

- Consider consulting community members and stakeholders about what racial categories best represent them and how erasure can manifest with fewer racial categories.
- Consider the epistemic goal of the dataset or model and how choices in racial categories contribute to this goal.
- Consider how the lived experiences of the dataset or model developers and researchers contribute to which racial categories are chosen.
- When developing a dataset, consider what racial categories of annotators and examples in the dataset are relevant.

*Documentation questions*

(1) What are the cultural backgrounds and cultural knowledge of the dataset or model developers? How familiar and/or knowledgeable are they with the cultural context(s) of the dataset or model they are developing?
(2) If CIRCSheets is completed by people other than the original dataset or model developers, what are their cultural backgrounds? How familiar and/or knowledgeable are they with the dataset or model's cultural context(s)?
(3) If annotators or crowd workers are used to develop a dataset or provide feedback to a model, what are their cultural backgrounds? How familiar and/or knowledgeable are they with the cultural context(s) of the instances they annotate?
(4) What stakeholders, community members, or other resources were consulted when selecting the racial categories?

# 5 CASE STUDY AND DISCUSSION

To demonstrate CIRCSheets in action, we apply our framework to the dataset associated with COMPAS using existing knowledge available about these datasets [4, 38].

## 5.1 Case Study: COMPAS

| Categories |
|---|

**What are the racial categories utilized?**
African-American, Asian, Caucasian, Hispanic, Native American, and Other.

**What is the motivation behind using these racial categories?**
No motivation is provided, but these categories seem to be taken from the US Census [29].

**Are multiracial ethnic categories utilized?**
Yes, Hispanic, a multiracial ethnicity, is treated as a race.

**If multiracial ethnic categories are used, what is the motivation behind using these categories, and are they being treated as a racial category?**
No motivation is provided by the dataset developers.

**Are people who select multiple racial categories considered multiracial? Are people who select one or more ethnic categories and one racial category considered multiracial?**
It is unclear if people can select multiple racial and/or ethnic categories, but in the dataset, each instance is assigned one racial or ethnic category. It is unclear whether people who select Hispanic and another race are considered multiracial. This is not discussed by the dataset developers [29].

**If so, what category are they placed into, and are other people who select multiple different racial and/or ethnic categories also placed into that same category? If not, why**

**category are they placed in, and does ethnicity take precedence over race?**
It is unclear what happens if people select multiple racial categories. It is possible that only one racial category is chosen for the individual from the ones they selected, or they are automatically placed into the "Other" category. The dataset developers do not discuss how multiracial individuals are categorized.

**For models, what is the technical implementation of the racial and/or ethnic categories?**
This is not applicable, as a model is not being used.

**How do ethnic groups fit into these racial categories?**
The dataset developers do not discuss this, but it seems that it follows the US Census with people who are descendants from the Black ethnicities of Africa are placed into the African-American category, people who are descendants of Asian ethnicities are placed into the Asian category, people who are descendants from European ethnicities are placed into the Caucasian category, and people with Native American ancestry are placed into the Native American category [47]. It seems that people with ancestry in Hispanic countries are placed into the Hispanic category, but it is not clear under what circumstances someone is placed into the Hispanic category rather than another racial category.

**Can people be obfuscated by these racial categories? If so, do these groups experience erasure, and is the model or dataset likely to interact with them?**
Yes. Because this dataset is centered in the US cultural context MENA (Middle Eastern and North African) individuals are most likely to be racialized as Caucasian. As discussed in Section 2.1, the experiences of MENA differ from the experiences of white people in the US [46]. Thus, it would not be possible to examine racial bias against MENA within COMPAS. There is also no category for Pacific Islanders, so it seems that those who identify as Pacific Islander would be placed into the "Other" racial category, which would obfuscate the experiences of Pacific Islanders.

"Other" has been used as a proxy for Latinx in the past [67]. It is unclear whether most individuals placed in the "Other" category in COMPAS identify with a particular ethnicity as was the case in the 1990 and 2000 US Censuses [67]. If "Other" can be used as a proxy for a particular ethnicity, this would further obfuscate groups, such as Pacific Islanders, who would be placed into the "Other" category.

## Racialization

**Who determines an individual's racial categorization? Is it the individual?**
It is unclear since the dataset developers do not discuss how a person's racial identity is determined [4, 29]

**Are physical characteristics asked of an individual?**
The dataset does not document physical characteristics, although race and gender are recorded. It is unclear if physical characteristics were asked of individuals to racialize them into a particular racial category.

**Is cultural background asked of an individual?**
It is not documented in the dataset.

**In what ways could the existing racial information be partial or incorrect? What impact could this have on the dataset or model?**
It is possible that some people racialized into the "Hispanic" or "Other" category were incorrectly racialized and should have been placed into another category. It is possible the features used to racialize people into categories were irrelevant to this particular domain.

This could impact the dataset because the dataset could have incorrect information, which would affect models trained on the dataset. These models may learn incorrect associations that, if deployed, would negatively impact the people affected by the model's decision. Furthermore, if the dataset is partially incorrect, auditing models would be more challenging since it would be unclear what information within the dataset is useful and what is irrelevant.

**If using an existing dataset and no racialization information exists, what was the source of the dataset, what cultural context was it developed in, and is there any existing scholarship on the racialization choices of that dataset?**
The dataset was developed in Broward County, Florida within the US cultural context. Existing scholarship on COMPAS discusses how "we don't know why the data take on a particular racial schema, nor do we have information about how defendants are racially categorized" [29, pp. 502]. Hanna et al. [29] discuss how the racial category an individual is placed into can change within a police department, so it is unclear how accurate the racial categories in COMPAS are for each individual even if the racial categorization schema were clearly communicated.

## Cultural Context

**What cultural context(s) is this dataset or model developed for?**
This dataset was developed in the US cultural context because it was developed in Broward County, Florida [29].

**Will this dataset or model be used in different cultural context(s)?**
It is possible this data may be used in different cultural contexts, but it seems unlikely as the dataset was created using US police records.

**If the dataset or model is used in different cultural context(s) and/or domains, is there any misrepresentation that can occur due to changes in racialization and/or racial categories in different cultural contexts and domains?**
Misrepresentation can occur if the dataset is used in different cultural contexts, as the racial categories seem chosen with the US cultural context in mind. Furthermore, it is unclear how these racial categories were developed and what aspects of racialization were most important in deciding what racial group people were placed into. This can become a greater issue if this dataset were used in a different cultural context. Furthermore, laws change depending on the country (and, in some cases, cities), so in different cultural

contexts, some people may not have been included in the dataset in the first place because their crime would not have been a crime in a different context

**Multiracial and Panethnicity**

**How are multiracial individuals and multiracial panethnicities categorized within the dataset or model?**
It is unclear how they are categorized within the dataset. It seems that only one category can be selected, so multiracial individuals may be placed in the "Other" racial category, or one of their racial identities may be chosen as their racial category. Either of these choices can have downstream impacts because the experiences of these multiracial individuals placed into these categories may differ from other individuals within this category.

The only multiracial panethnicity considered in the dataset is Hispanic, and it is treated as a race. It is unclear how Black and white Hispanics would be categorized. Any categorization schema based on the singular racial categories provided could obfuscate identities. If the Hispanic category supersedes the African-American or Caucasian category, then the experiences of Black Hispanics would be obfuscated. If race supersedes, then the experiences of both Black and white Hispanics would be obfuscated by the racial categories they have been placed in since their experiences differ from other Black and white individuals.

**Can more than one racial category be selected?**
No.

**Do the categories given to panethnic individuals effectively communicate their racial and ethnic identities?**
No, because only one category can be selected.

**Are there any individuals, such as Afro-Latinxs, who would not be adequately represented by the racial categorizations chosen?**
Yes, any multiracial individual or any individual who is racialized outside of their panethnicity, like Afro-Latinxs.

**Knowledge and Positionality**

**What are the cultural backgrounds and cultural knowledge of the dataset or model developers? How familiar and/or knowledgeable are they with the cultural context(s) of the dataset or model they are developing?**
This is unknown as no information was released from Broward County, Florida regarding this.

**If CIRCSheets is filled out by people other than the original dataset or model developers, what are their cultural backgrounds? How familiar and/or knowledgeable are they with the cultural context(s) of the dataset or model?**
The individual filling this out is a Russian-American woman who grew up in the US cultural context, so she is familiar with US racial structures.

**If annotators or crowd workers are used, what are their cultural backgrounds? How familiar and/or knowledgeable are they with the cultural context(s) of the instances they annotate?**
This is unknown as no information was released about this from Broward County, Florida.

**What stakeholders, community members, or other resources were consulted when deciding the racial categories?**
This is unknown as no information was released about this from Broward County, Florida.

## 5.2 Discussion

The case study presented above illustrates a completed CIRCSheet for COMPAS. As this is a pre-existing dataset, not all of the questions could be answered, as some information was unavailable. Nevertheless, users of COMPAS could utilize this CIRCSheet to improve their understanding of how people are racialized into the racial groups used in COMPAS. CIRCSheets is beneficial as it helps practitioners understand the racialization processes used to develop datasets and models. Practitioners and researchers utilizing an existing CIRCSheet for a dataset or model would have a better understanding of how the racial categories were selected and who is racialized into them.

For practitioners and researchers completing a CIRCSheet for a dataset or model, the practice of going through the questions and answering them would assist practitioners and researchers in considering why they selected particular racial and ethnic categories and how people are racialized into those categories. Furthermore, they would consider how their choice of racial schema may lead to the obfuscation of people with certain racial or ethnic identities. Ananny and Crawford [2] discuss the limitations of transparency, but with CIRCSheets, the process of considering *why* particular racial and ethnic categories are selected, and the consequences of this selection process is a primary benefit of this documentation framework, as it can lead to more meaningful racial and ethnic categories. Furthermore, it is crucial for practitioners and researchers who utilize CIRCSheets to critically engage with the questions and consider the implications of their choices. The benefits of CIRCSheets would lessen if practitioners and researchers treated CIRCSheets as an additional checklist item and did not critically engage with the questions.

# 6 CONCLUSION

In this work, we discuss the importance of racial and ethnic categories and demonstrate the effect these choices can have on dataset quality and model performance with different interpretations of racial categories and racialization processes. Therefore, to facilitate understanding of the racial categories and racialization processes used, we develop CIRCSheets as a documentation tool for developers to communicate their assumptions, motivations, and racialization understanding, as well as, potential pitfalls. This documentation allows future users to better understand the racial and ethnic categories documented and how people are placed into these categories, assisting them in determining whether they can use this information in future tasks, such as auditing datasets and models

or deploying models to consumers. Dataset and model users can also use CIRCSheets to communicate their own understanding of existing racial categories when information regarding the racial categories and racialization process in existing datasets and models is unclear or does not exist.

## ACKNOWLEDGMENTS

Thank you to Yasmiyn Irizarry, Kevin Tian, and four anonymous FAccT reviewers for helpful comments on earlier versions of this paper. Thank you to the students of Dr. Irizarry's Theories of Race and Ethnicity class for the insightful conversations that shaped this work.

## REFERENCES

[1] Amina A Abdu, Irene V Pasquetto, and Abigail Z Jacobs. 2023. An Empirical Analysis of Racial Categories in the Algorithmic Fairness Literature. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1324–1333.

[2] Mike Ananny and Kate Crawford. 2018. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *new media & society* 20, 3 (2018), 973–989.

[3] McKane Andrus and Sarah Villeneuve. 2022. Demographic-reliant algorithmic fairness: Characterizing the risks of demographic data collection in the pursuit of fairness. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1709–1721.

[4] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2022. Machine bias. In *Ethics of data and analytics*. Auerbach Publications, 254–264.

[5] Germine H Awad, Nadia N Abuelezam, Kristine J Ajrouch, and Matthew Jaber Stiffler. 2022. Lack of Arab or Middle Eastern and North African health data undermines assessment of health disparities. *American Journal of Public Health* 112, 2 (2022), 209–212.

[6] Darwin A Baluran. 2023. Life expectancy, life disparity, and differential racialization among Chinese, Asian Indians, and Filipinos in the United States. *SSM-Population Health* 21 (2023), 101306.

[7] Steven D Barger, Carrie J Donoho, and Heidi A Wayment. 2009. The relative contributions of race/ethnicity, socioeconomic status, health, and social relationships to life satisfaction in the United States. *Quality of Life Research* 18 (2009), 179–189.

[8] Vilna Bashi and Antonio McDaniel. 1997. A theory of immigration and racial stratification. *Journal of Black Studies* 27, 5 (1997), 668–682.

[9] Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604.

[10] Sebastian Benthall and Bruce D Haynes. 2019. Racial categories in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*. 289–298.

[11] Maximus Berger and Zoltán Sarnyai. 2015. "More than skin deep": stress neurobiology and mental health consequences of racial discrimination. *Stress* 18, 1 (2015), 1–10.

[12] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. Fairlearn: A toolkit for assessing and improving fairness in AI. *Microsoft, Tech. Rep. MSR-TR-2020-32* (2020).

[13] Eduardo Bonilla-Silva. 2004. From bi-racial to tri-racial: Towards a new system of racial stratification in the USA. *Ethnic and racial studies* 27, 6 (2004), 931–950.

[14] Glencora Borradaile, Brett Burkhardt, and Alexandria LeClerc. 2020. Whose tweets are surveilled for the police: an audit of a social-media monitoring tool via log files. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 570–580.

[15] Robert T Carter, Michael Y Lau, Veronica Johnson, and Katherine Kirkinis. 2017. Racial discrimination and health outcomes among racial/ethnic minorities: A meta-analytic review. *Journal of Multicultural Counseling and Development* 45, 4 (2017), 232–259.

[16] Kim D Chanbonpin. 2015. Between black and white: The coloring of Asian Americans. *Wash. U. Global Stud. L. Rev.* 14 (2015), 637.

[17] Kasia S Chmielinski, Sarah Newman, Matt Taylor, Josh Joseph, Kemi Thomas, Jessica Yurkofsky, and Yue Chelsea Qiu. 2022. The dataset nutrition label (2nd Gen): Leveraging context to mitigate harms in artificial intelligence. *arXiv preprint arXiv:2201.03954* (2022).

[18] Matthew Clair and Jeffrey S Denis. 2015. Sociology of racism. *The international encyclopedia of the social and behavioral sciences* 19, 2015 (2015), 857–63.

[19] Stephen Cornell and Douglas Hartmann. 2006. *Ethnicity and race: Making identities in a changing world*. Sage Publications.

[20] Anamaria Crisan, Margaret Drouhard, Jesse Vig, and Nazneen Rajani. 2022. Interactive model cards: A human-centered approach to model documentation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 427–439.

[21] F James Davis. 2010. *Who is black?: One nation's definition*. Penn State Press.

[22] Michelle Daya, Lize Van Der Merwe, Ushma Galal, Marlo Möller, Muneeb Salie, Emile R Chimusa, Joshua M Galanter, Paul D Van Helden, Brenna M Henn, Chris R Gignoux, et al. 2013. A panel of ancestry informative markers for the complex five-way admixed South African coloured population. *PloS one* 8, 12 (2013), e82224.

[23] Mark Díaz, Ian Kivlichan, Rachel Rosen, Dylan Baker, Razvan Amironesei, Vinodkumar Prabhakaran, and Emily Denton. 2022. Crowdworksheets: Accounting for individual and collective identities underlying crowdsourced dataset annotation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 2342–2351.

[24] Špela Drnovšek Zorko and Miloš Debnár. 2021. Comparing the racialization of Central-East European migrants in Japan and the UK. *Comparative Migration Studies* 9, 1 (2021), 1–17.

[25] Karen Farquharson. 2007. Racial categories in three nations: Australia, South Africa and the United States. In *Proceedings of 'Public sociologies: lessons and trans-Tasman Comparisons', the Annual Conference of The Australian Sociological Association (TASA)*.

[26] Karly S Ford, Ashley N Patterson, and Marc P Johnston-Guerrero. 2021. Monoracial normativity in university websites: Systematic erasure and selective reclassification of multiracial students. *Journal of Diversity in Higher Education* 14, 2 (2021), 252.

[27] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*. 329–338.

[28] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.

[29] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 501–512.

[30] Tanya Katerí Hernández. 2022. *Racial innocence: Unmasking Latino anti-Black bias and the struggle for equality*. Beacon Press.

[31] Jennifer L Hochschild and Brenna Marea Powell. 2008. Racial reorganization and the United States Census 1850–1930: Mulattoes, half-breeds, mixed parentage, Hindoos, and the Mexican race. *Studies in American Political Development* 22, 1 (2008), 59–96.

[32] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2020. The dataset nutrition label. *Data Protection and Privacy* 12, 12 (2020), 1.

[33] Elizabeth Hordge-Freeman and Edlin Veras. 2020. Out of the shadows, into the dark: Ethnoracial dissonance and identity formation among Afro-Latinxs. *Sociology of Race and Ethnicity* 6, 2 (2020), 146–160.

[34] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 560–575.

[35] Yasmiyn Irizarry. 2015. Utilizing multidimensional measures of race in education research: The case of teacher perceptions. *Sociology of Race and Ethnicity* 1, 4 (2015), 564–583.

[36] Yasmiyn Irizarry, Ellis P Monk Jr, and Ryon J Cobb. 2023. Race-shifting in the United States: Latinxs, Skin Tone, and Ethnoracial Alignments. *Sociology of Race and Ethnicity* 9, 1 (2023), 37–55.

[37] Jose Itzigsohn, Silvia Giorguli, and Obed Vazquez. 2005. Immigrant incorporation and racial identity: Racial self-identification among Dominican immigrants. *Ethnic and Racial Studies* 28, 1 (2005), 50–78.

[38] Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, et al. 2022. Introducing the Gab Hate Corpus: defining and applying hate-based rhetoric to social media posts at scale. *Language Resources and Evaluation* (2022), 1–30.

[39] Vijay Keswani, Matthew Lease, and Krishnaram Kenthapadi. 2021. Towards unbiased and accurate deferral to multiple experts. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 154–165.

[40] Zaid Khan and Yun Fu. 2021. One label, one billion faces: Usage and consistency of racial categories in computer vision. In *Proceedings of the 2021 acm conference on fairness, accountability, and transparency*. 587–597.

[41] Nikki Khanna. 2010. "If you're half black, you're just black": Reflected appraisals and the persistence of the one-drop rule. *The Sociological Quarterly* 51, 1 (2010),

96–121.
[42] Michelle Seng Ah Lee and Jat Singh. 2021. The landscape and gaps in open source fairness toolkits. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–13.
[43] Sharon M Lee. 1993. Racial classifications in the US Census: 1890–1990. *Ethnic and racial studies* 16, 1 (1993), 75–94.
[44] Ian Haney Lopez. 2006. White by law 10th anniversary edition: The legal construction of race. In *White by Law 10th Anniversary Edition*. New York University Press.
[45] Nancy López. 2013. Killing two birds with one stone? Why we need two separate questions on race and ethnicity in the 2020 census and beyond. *Latino Studies* 11 (2013), 428–438.
[46] Neda Maghbouleh, Ariela Schachter, and René D Flores. 2022. Middle Eastern and North African Americans may not be perceived, nor perceive themselves, to be White. *Proceedings of the National Academy of Sciences* 119, 7 (2022), e2117940119.
[47] Rachel Marks and Nicholas Jones. 2020. Collecting and Tabulating Ethnicity and Race Responses in the 2020 Census. *United States Census Bureau* (2020). https://www2.census.gov/about/training-workshops/2020/2020-02-19-pop-presentation.pdf
[48] Anthony W Marx. 1998. *Making race and nation: A comparison of South Africa, the United States, and Brazil*. Cambridge University Press.
[49] Kay Young McChesney. 2015. Teaching diversity: The science you need to know to explain why race is not biological. *SAGE Open* 5, 4 (2015), 2158244015611712.
[50] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.
[51] Ellis P Monk Jr, Michael H Esposito, and Hedwig Lee. 2021. Beholding inequality: Race, gender, and returns to physical attractiveness in the United States. *Amer. J. Sociology* 127, 1 (2021), 194–241.
[52] G Cristina Mora, Reuben Perez, and Nicholas Vargas. 2022. Who identifies as "Latinx"? The generational politics of ethnoracial labels. *Social Forces* 100, 3 (2022), 1170–1194.
[53] Carol C Mukhopadhyay, Rosemary Henze, and Yolanda T Moses. 2013. *How real is race?: A sourcebook on race, culture, and biology*. Rowman & Littlefield.
[54] Karim Murji and John Solomos. 2005. *Racialization: Studies in theory and practice*. Oxford University Press, USA.
[55] Laura Beth Nielsen. 2002. Subtle, pervasive, harmful: Racist and sexist remarks in public as hate speech. *Journal of Social issues* 58, 2 (2002), 265–280.
[56] Suryadewi E Nugraheni and Julia F Hastings. 2021. Family-based caregiving: Does lumping Asian Americans together do more harm than good? *Journal of Social, Behavioral, and Health Sciences* 15, 1 (2021), 87–102.
[57] Anthony C Ocampo. 2013. " Am I Really Asian?": Educational Experiences and Panethnic Identification among Second–Generation Filipino Americans. *Journal of Asian American Studies* 16, 3 (2013), 295–324.
[58] Anthony Christian Ocampo. 2016. *The Latinos of Asia: How Filipino Americans break the rules of race*. Stanford University Press.
[59] Dina Okamoto and G. Cristina Mora. 2014. Panethnicity. *Annual Review of Sociology* 40, 1 (2014), 219–239. https://doi.org/10.1146/annurev-soc-071913-043201 arXiv:https://doi.org/10.1146/annurev-soc-071913-043201
[60] Olihe N Okoro, Vibhuti Arya, Caroline A Gaither, and Adati Tarfa. 2021. Examining the inclusion of race and ethnicity in patient cases. *American journal of pharmaceutical education* 85, 9 (2021), 8583.
[61] Eric Steven O'Malley. 2000. Irreconcilable rights and the question of Hawaiian statehood. *Geo. LJ* 89 (2000), 501.
[62] Michael Omi and Howard Winant. 2014. *Racial formation in the United States*. Routledge.
[63] Whitney N Laster Pirtle. 2021. Racial states and re-making race: Exploring coloured racial re-and de-formation in state laws and forms in post-apartheid South Africa. *Sociology of Race and Ethnicity* 7, 2 (2021), 145–159.
[64] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. 2022. Data cards: Purposeful and transparent dataset documentation for responsible ai. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1776–1826.
[65] Timothy R Rebbeck, Brandon Mahal, Kara N Maxwell, Isla P Garraway, and Kosj Yamoah. 2022. The distinct impacts of race and genetic ancestry on health. *Nature medicine* 28, 5 (2022), 890–893.
[66] Jessica D Remedios and Alison L Chasteen. 2013. Finally, someone who "gets" me! Multiracial people value others' accuracy about their race. *Cultural Diversity and Ethnic Minority Psychology* 19, 4 (2013), 453.
[67] Clara E Rodriguez. 2000. *Changing race: Latinos, the census, and the history of ethnicity in the United States*. Vol. 41. NYU Press.
[68] Wendy D Roth. 2010. Racial mismatch: The divergence between form and function in data for monitoring racial discrimination of Hispanics. *Social Science Quarterly* 91, 5 (2010), 1288–1311.
[69] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry* 22, 2014 (2014), 4349–4357.
[70] Paul Schor. 2017. *Counting Americans : how the US Census classified the nation*. Oxford University Press, New York, NY.
[71] Gail Shuck. 2006. Racializing the nonnative English speaker. *Journal of Language, Identity, and Education* 5, 4 (2006), 259–276.
[72] Edward Telles. 2018. Latinos, race, and the US census. *The ANNALS of the American Academy of Political and Social Science* 677, 1 (2018), 153–164.
[73] Edward Telles and Tianna Paschel. 2014. Who is black, white, or mixed race? How skin color, status, and nation shape racial classification in Latin America. *Amer. J. Sociology* 120, 3 (2014), 864–907.
[74] Fernando M Treviño. 1987. Standardized terminology for hispanic populations. *American Journal of Public Health* 77, 1 (1987), 69–72.
[75] Ekeoma E Uzogara. 2021. Who belongs in America? Latinxs' skin tones, perceived discrimination, and opposition to multicultural policies. *Cultural diversity and ethnic minority psychology* 27, 3 (2021), 354.
[76] Briana Vecchione, Karen Levy, and Solon Barocas. 2021. Algorithmic auditing and social justice: Lessons from the history of audit studies. In *Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–9.
[77] Sarahn M Wheeler and Allison S Bryant. 2017. Racial and ethnic disparities in health and health care. *Obstetrics and Gynecology Clinics* 44, 1 (2017), 1–11.
[78] Robert Wolfe, Mahzarin R Banaji, and Aylin Caliskan. 2022. Evidence for hypodescent in visual semantic AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1293–1304.
[79] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. 2020. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 547–558.
[80] Henry Yu. 2020. *27 Ethnicity*. New York University Press, New York, USA, 106–110. https://doi.org/doi:10.18574/nyu/9781479867455.003.0031