How We've Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis

MORGAN KLAUS SCHEUERMAN, University of Colorado Boulder, USA
KANDREA WADE, University of Colorado Boulder, USA
CAITLIN LUSTIG, University of Washington, USA
JED R. BRUBAKER, University of Colorado Boulder, USA

Race and gender have long sociopolitical histories of classification in technical infrastructures—from the passport to social media. Facial analysis technologies are particularly pertinent to understanding how identity is operationalized in new technical systems. What facial analysis technologies can do is determined by the data available to train and evaluate them with. In this study, we specifically focus on this data by examining how race and gender are defined and annotated in image databases used for facial analysis. We found that the majority of image databases rarely contain underlying source material for how those identities are defined. Further, when they are annotated with race and gender information, database authors rarely describe the process of annotation. Instead, classifications of race and gender are portrayed as insignificant, indisputable, and apolitical. We discuss the limitations of these approaches given the sociohistorical nature of race and gender. We posit that the lack of critical engagement with this nature renders databases opaque and less trustworthy. We conclude by encouraging database authors to address both the histories of classification inherently embedded into race and gender, as well as their positionality in embedding such classifications.

CCS Concepts: • **Social and professional topics → User characteristics**; **Race and ethnicity**; **Gender**.

Additional Key Words and Phrases: Classification; computer vision; facial analysis; facial classification; facial recognition; facial detection; training and evaluation data; identity; gender; race and ethnicity

ACM Reference Format:
Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R. Brubaker. 2020. How We've Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1, Article 58 (May 2020), 35 pages. https://doi.org/10.1145/3392866

1 Introduction

Race and gender are a part of our everyday reality. We all experience the visible external features associated with race and gender when we interact with the world. We also hold our own internal and otherwise invisible affinities with these identities. Race and gender have also always been incorporated into technical systems. From the U.S. Census [164] to Facebook [20], we see technical representations of these identities everywhere. They have also now become commonly embedded into databases as features used to train new machine-learning based algorithms. Facial analysis

Authors' addresses: Morgan Klaus Scheuerman, University of Colorado Boulder, Department of Information Science, Boulder, CO, 80309, USA, morgan.scheuerman@colorado.edu; Kandrea Wade, University of Colorado Boulder, Department of Information Science, Boulder, CO, 80309, USA, kandrea.wade@colorado.edu; Caitlin Lustig, University of Washington, Human Centered Design and Engineering, Seattle, WA, 98195, USA, celustig@uw.edu; Jed R. Brubaker, University of Colorado Boulder, Department of Information Science, Boulder, CO, 80309, USA, jed.brubaker@colorado.edu.

2573-0142/2020/5-ART58 $15.00
https://doi.org/10.1145/3392866

technology—the machine learning (ML) approach to determining information about human faces—is just the latest tool in a long history of tools used for classifying human identity.

Within the last decade, facial analysis technology has quickly transitioned from a theoretical research problem to commercial reality. Facial recognition is already embedded in technologies (e.g., social media [119], device locking [5]) and employed by law enforcement ([53, 165]. Facial classification is used to target marketing campaigns at specific demographics (e.g., [128, 139, 154]) and to track physical consumer behavior inside stores (e.g., [4, 70, 96]). The outputs produced by facial analysis systems are premised on their training and evaluation data: the images used to "teach" the system what a subject looks like.

The use of human identity characteristics as outputs have been increasingly scrutinized by computing researchers (e.g., [80, 134, 150]). Race and gender have become two of the largest concerns regarding bias in machine learning fairness literature—particularly, how systems are biased against certain races and genders [6, 58, 84] and how to mitigate those biases [24, 55, 167]. These concerns include bias in the databases used to train and evaluate machine learning algorithms (e.g., [34, 108, 161]) and the very morality of facial analysis use cases (e.g., [14, 104, 169]). Sample selection bias–bias resulting from what subjects are included in a database–is a known issue in machine learning databases (e.g., [108, 162]), leading computer scientists to try to mitigate for it using various methods. For example, algorithms for exploiting database bias to improve those databases (e.g., [83]) and for creating "unbiased" models from known biased data (e.g., [76]).

Such attempts to mitigate bias and build more diverse databases are invaluable to creating fairer outcomes. However, despite increasing attempts to diversify databases, approaches remain simplistic and lacking in critical and social theories. ML and human-computer interaction (HCI) communities do not have an agreed upon approach to how diversity is being operationalized in training and evaluation databases. New databases are seeking to fill gaps with more images without a deeper engagement of the categories of race and gender themselves or the ethics of collecting that information (e.g., Google contractors targeting homeless people of color for face images [67]). While some scholars are questioning the politics of identity representations in facial analysis classification infrastructures (e.g., [80, 150]), there has been little inquiry into the assumptions authors of facial analysis databases have made when collecting and annotating data. This is a major obstacle in meaningfully representing race and gender in databases, resulting in databases that are opaque and inconsistent. To truly understand the available outcomes of facial analysis models, it is imperative to understand the underlying decisions embedded into the construction of training and evaluation databases.

Like Benthall and Hayes in [18], we examine race—and gender—as socially constructed categories machine learning has failed to critically engage with. We approach our analysis from a critical discursive perspective. Specifically, we investigate *how* race and gender are codified into image databases. To do this, we analyze how race and gender are represented in image databases and how those representations are derived. We focus on answering the following research questions:

(1) What *purposes* do the authors of image databases intend their databases to be used for and how does that shape their use of race and gender?
(2) What information about race and/or gender are *implicit* (i.e., the authors describe the demographic distribution in a database, but each image is not annotated with race and/or gender) and what information is *explicit* (i.e., each image in a database is annotated with race and/or gender information)? What are the *categories* being used to define race and gender?
(3) What *sources* are being used to derive race and/or gender in both implicit and explicit databases?

(4) How are database authors describing the *annotation procedures* for explicitly annotated race and/or gender categories?

To identify relevant databases for analysis, we created a corpus of machine learning literature on facial analysis technologies and manually coded them for which databases are referenced. We used this corpus to identify a sample of 92 image databases, whose documentation we examined to answer the outlined research questions. We started by analyzing the database documentation to identify the motivations authors provided for the use of each database—in other words, what each database was created for. Understanding these motivations provided context for the intended uses of race and gender in facial analysis systems. We found three database use cases: (1) individual face recognition and verification; (2) image labeling and classification; and (3) providing diversity for model training and evaluation. We then surveyed both (1) implicit race and gender information; and (2) explicit race and gender annotations. We chose to analyze both implicit and explicit descriptions as both have implications for database use, value, and potential bias.

Within both implicit and explicit race and gender categories, we found two diverging themes. For race, we observed no consistent classification schema; the classification of race and the way race is discussed by authors varies greatly. For gender, like previous authors [80, 150], we observed numerous instantiations of the same "male" and "female" binary categories. We found that the vast majority of image databases (1) do not utilize sources (e.g., make use of existing resources, like prior literature) for defining race and gender categories; and (2) do not document the process of annotating images for race and gender categories.

Given that image databases are used as a resource on which facial analysis systems are built and evaluated, we argue that the field of computer vision needs to adopt more standardized methods for using and documenting race and gender. We posit facial analysis as a digital form of otherwise familiar classification technologies to critique current approaches in image databases for their lack of critical engagement with racial and gender histories. We discuss race and gender categories in technical databases through a multidisciplinary lens, synthesizing theory from critical race studies, gender studies, infrastructural studies, and identity scholarship. We build on previous fairness scholarship to specify options for the field of computer vision, and machine learning more broadly, to evolve its approach to human identity and embrace new lines of research. Our findings highlight opportunities for more human-centered methods that will improve both the representation of race and gender and the validity of annotations in image databases.

2 Related Work

We situate our work at the intersection of two major areas of scholarship: identity theory and machine learning. Within these two areas, we weave together work from multiple disciplines to frame the complicated and nuanced reality of categorizing identity in facial analysis technologies.

We begin by discussing how complex the concept of identity is, highlighting numerous theories for defining identity in philosophy, gender studies, and critical race theory. We discuss how visible aspects of humans have been used to classify people into differing identity categories, drawing connections between historical practices of racial and gender identification with more current technical practices. Given the focus on the visual, we turn to the lens of "identification," as defined by Stuart Hall in [61], as a means of better interrogating current practices of facial analysis databases. Finally, we turn to recent work in machine learning on fairness, accountability, transparency, and ethics (FATE) that has specifically focused on identity inequity. We spotlight current work aimed at uncovering and mitigating both racial and gender biases in machine learning.

2.1 (In)visible Characteristics: Human Identity in Theory and Practice

The concept of identity, from human to computer, has erupted as a site of inquiry with the rise of disciplines like HCI, society and technology studies (STS), and digital humanities. Technology scholars have explored concepts such as embodied identity in virtual worlds [10, 21, 71], including intersecting experiences of race and gender [117]. They have also centered human identities in their studies of social media platforms, uncovering, for example, the emotional labor of moderators on Asian American and Pacific Islander identity forums [38]; the benefits of online fandom communities in LGBTQ+ identity exploration [40]; and how the systematic discrimination against trans people can bridge both online and offline spaces [149].

Identity, broadly, is often divided into two perspectives: the visible and the invisible. Locke argued that personal sense of self might belong to the "consciousness," having nothing to do with visible physical embodiment [99]. In contrast, scholars such as Husserl, Butler, and Alcoff assert that the embodied self is central to the development of internal awareness and one's relationship with the world [9, 25, 110]. Alcoff, in particular, embraces notions of visual embodiment in [9], discussing the significance of visible markers of identity for race and gender in discussions of social identity, particularly in opposition to perspectives that seek to erase race and gender from political discussions. Namaste similarly critiques the overly philosophized perspective on gender, centering trans people's lived and embodied experiences which are often erased, or made invisible, from gendered theory [118].

However, classifying humans based on visible difference has sometimes been utilized for technologies of oppression. The practice of physiognomy, the notion that certain races' bodily and facial structures communicate inferior internal mental and emotional qualities, has been observed in both the Rwandan genocide [86] and the long history of slavery and racial segregation in the United States [48, 156]. Technology researchers have begun drawing lines between these histories and technological practices of identity representation. Phillips connected past physiognomic practices to shaping racial models in video games [127]. Benthall and Haynes questioned the racial categories employed in broader machine learning practices, particularly in how they've been shaped by histories of political oppression against Black people [18]. Bowker and Star outline numerous examples of how technologies that were deployed for racial segregation manifested in Apartheid South Africa in [22], highlighting the evolving process of classification and reclassification that shaped and twisted people's lives. Some commercial facial analysis companies have also adopted notions of physiognomy, such as Faception which attempts to tell internal characteristics, like IQ and criminality, from facial morphology [107].

Institutional classifications of identity are constantly shifting; they are not solely remnants of the past. For example, in the United States, President Trump recently issued an executive order to extend racial and ethnic classification to those with Jewish ancestry in an effort to curb anti-Israel protests on college campuses under Title IX [36]. Furthermore, classifications systems are often implemented unevenly. While some U.S. states allow for non-binary genders on birth certificates (e.g., [148]) and driver's licenses (e.g., [33, 42]), it remains impossible to change gender markers in other states. The debates and laws around gender classifications are still evolving. Many trans people have criticized restrictive gender classification due to the increased likelihood trans people will be further exposed to risky interactions with police and security officials [62, 106, 116]. Methods of identifying and tracking marginalized people have already extended to new digital tools. Today, we see facial analysis being deployed in similarly oppressive ways. In China, Uyghur Muslim minorities, who are increasingly being detained in re-education camps, are subject to government surveillance by facial classification and recognition, trained explicitly to attempt to classify and track people who appear Uyghur [113].

Stuart Hall gave us the lens of *identification* [61] through which to problematize identity classifications. To Hall, identification signifies the "process of articulation, a suturing" [61]. It can mean an identification *with* (for example, a shared history) or an identification *of* (for example, an assigned sex [25]). He describes the construction of identity categories within "specific modalities of power," for which identities like race and gender are employed for discursive means. The lens of identification opens interesting avenues for viewing perspectives of identity as it pertains to digital technologies. Specifically, how technologies are designed with the identification of human identity features in mind. In this paper, we use identification to interrogate how database authors use visible information about subjects to make determinations about race and gender. We argue that there are historical and theoretical limitations to current database annotation approaches in machine learning research. Through an examination of our findings, we discuss the potential for historic identification biases to become embedded at the first step of machine learning model building: the data it is limited to learning from.

2.2 FATE ML: Fairness, Accountability, Transparency, and Ethics in Machine Learning

Bias in technical systems results in unfair outcomes for differing people—often, for those of marginalized race and gender identities. To address this, bias auditing and bias mitigation have become a major focus of both computer science and HCI research. The rise of FATE (fairness, accountability, transparency, and ethics) for AI development has yielded new insights, methods, and frameworks for addressing issues of bias. Bias has been conceptualized in various ways by researchers—for example, statistical bias that results in skewed results [35, 72] and representation bias stemming from historical prejudice [69, 108]. These efforts have been fused with scholarship from critical algorithm studies focused on identity. This is evident in Noble's analysis of Black women's representations in Google search results [122] and Hamilton's findings of the overzealous risk scoring of women in recidivism risk algorithm, COMPAS [63]. Similarly, Obermeyer et al. uncovered racial bias against Black patients in commercial health risk prediction algorithms, linked to historic data about medical spending [123]; and Kay et al. found gender stereotypes about women's occupational roles in image search results [79].

Bias in algorithmic contexts can cause widespread, real world harm. The Future Privacy Forum released a report categorizing the numerous harms that can result from algorithmic bias, grouped by individual-level harms (e.g., employment discrimination) and societal-level harms (e.g., differential access to job opportunities) [3]. Given the reality of algorithmic bias and its inevitable consequences, scholars across disciplines have sought solutions. Machine learning researchers have proposed numerous statistical approaches (e.g., [35]) and toolkits (e.g., [17]) for mitigating bias. As bias can manifest in numerous ways and in numerous places within a machine learning system, many scholars have began considering its consequences. For example, Danks and London present a taxonomy of where algorithmic bias might appear in the pipeline in [34]: in the training data, in the focus of the algorithm, in the processing of information, and in the use of a single algorithm from one context to another. Hanna et al. proposed a critical race methodology for machine learning, to better adapt to the complex realities of race [64].

Already, much research has been done to uncover bias in facial analysis systems. NIST conducted an evaluation of face recognition in 2019, finding that recognition systems tend to perform better on men and older people, than on women and younger people [58]. Klare et al. similarly discovered that models performed worse on women, as well as people who are Black [84]. Buolamwini and Gebru found that facial analysis services, like those provided by Microsoft and IBM, had significantly higher gender misclassification rates for women with dark skin tones [24]. Further, scholars and practitioners are questioning the underlying social and moral judgments being made when detecting and classifying core human identities. Hamidi et al. interviewed transgender (i.e.,

trans) individuals and found their participants were largely concerned about how facial analysis could be used for discrimination [62]. Similarly, Scheuerman et al. found gender classification in facial analysis consistently performed worse on trans people in comparison with cis people, prompting a discussion on how gender is defined in gender classification models and what purpose it serves [150].

Our study embraces previous work's call for more nuanced examinations of fairness by starting at the source: the data. Specifically, we interrogate the notion of race and gender classifications in image databases as inherently and objectively classifiable, drawing connections between past technological oppression of race and gender. Through our approach, we encourage machine learning experts to consider critical theory, human-centered approaches, and technical histories when making present-day decisions about facial analysis tasks. We propose more rigorously documented decision-making processes when dealing with race and gender data.

3 Background Information on Image Databases for Facial Analysis Technologies

Automated facial analysis is a subset of computer vision technology designed for the specific task of analyzing human faces in digital images and videos. Facial analysis technologies are built using a number of machine learning approaches (e.g., [52, 95, 158]), generally to accomplish two goals: facial recognition and facial classification. Facial recognition, first developed in the 1960s, is designed to match an image to a specific individual [12]. Recognition is now used by social media platforms (e.g., Facebook's tagging system [119]), consumer electronics (e.g., iPhone's Apple ID [5]), and police departments (e.g., [165]). Facial classification is designed to label specific features about a human face. For example, the perceived gender (e.g., [136, 141, 147]) or ethnicity (e.g., [60, 100, 103]) of a face; whether the face is considered beautiful or not (e.g., [41, 170]); and even what the face can tell us about a person's criminality, sexuality, or intelligence (e.g., [107, 168, 173]). While facial recognition and facial classification are separate tasks, they are both posited on successful facial detection and feature extraction. The same software may be designed to complete both tasks (e.g., [101]). A simplified diagram of facial recognition and classification pipelines can be seen in Figure 1.

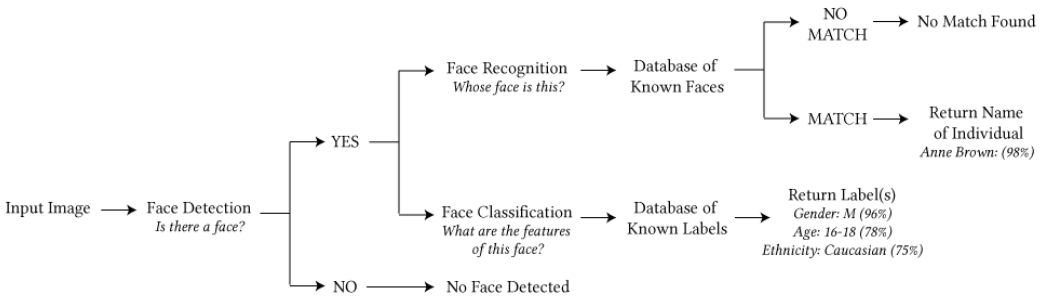

Fig. 1. A diagram of facial analysis tasks: one pipeline represents facial recognition, the other facial classification. The diagram represents one simple approach; there are numerous other approaches developers might take, including using facial classification to aid facial recognition (e.g., [101]).

Underlying how facial analysis systems work are the databases of images and/or videos that the systems are trained and evaluated on. Our sites of inquiry are the databases that enable the training and evaluation of facial analysis models. Before annotation is possible, there must be data

to be annotated. Facial data is can be collected in a number of ways. Early databases were often constructed by recruiting individuals and collecting multiple photographs of them in a studio (e.g., Yale Face Database [16], FERET [75]). As databases have grown larger to meet the demands of increasingly complex computer vision tasks, databases are now being built by scraping publicly available image data, such as the People In Photo Albums (PIPA) database, which used Flickr images [175]).

Posterior to data collection, the images are often annotated, or labelled, with specific information that are then useful to facial analysis. Databases may be annotated differently depending on the task the database was intended to be used for. For example, some databases may be annotated with identifiable information for each individual person in an image, such as a name. Another may be annotated with characteristic information about an individual, such as their race or gender. Annotation may be done via a number of methods. Smaller databases may be annotated by the original creators, as can be seen in databases like Pilot Parliaments Benchmark (PPB) [24]. Increasingly, databases are being annotated through crowdsourcing or scraping associated image information on the web [142].

We examined image databases that are then used to train and evaluate facial analysis technologies as a site of inquiry. Specifically, we examine the documentation of image databases to determine how race and gender is being accounted for. We specifically analyze how race and gender are determined and labelled by database authors. We discuss our approach to conducting this study in the next section.

4 Methods

4.1 Researcher Positionality

Reflexivity in research practice establishes the researcher as a lens through which research is conducted; what Attia and Edge call an "on-going mutual shaping between researcher and research" [13]. In other words, the research is shaped by the positionality—the social and political context—of the researcher. In alignment with the feminist practice of reflexively examining one's relationship to one's research, we want to highlight how the positionality of the authors may have shaped this work.

Our approaches to examining both gender and race are informed by our collective experience—and many other experiences that make up our perspectives as researchers. The second author is Black, while the remaining three authors are white. Every author has a different gender. All of the authors are based in the United States. As such, our experiences are rooted in a Western-centric point of view. Each author comes from a multidisciplinary background, including HCI, computer science, psychology, gender studies, communication, and the arts. Our synthesized experiences with critical theory, race studies, and gender studies are shaped by education (both formal and informal) related to our U.S. nationalities. Our decision to examine computer vision practice with a critical lens stems from our scholarly upbringings. Our privilege as academics awarded us access to the resources to conduct this work, while our power as differentially marginalized individuals gave us the perspective to develop our research questions and interpret our data.

4.2 Defining the Language of Analysis

Through our analysis, we discovered that different terms were being used interchangeably to describe race. This was true for gender as well. For that reason, we will outline how we have decided to define our use of the terms "race" and "gender" in the context of this paper. We return to practices of defining race and gender in more depth in our Design Considerations (see section 9).

4.2.1 *Defining "Race"*

Documentation of image databases often used both "ethnicity" and "race" to refer to a single concept: an annotation based on phenotypic attributes like skin tone and facial features perceived to be relevant to differing ancestral histories. Critical race theory and histories of racial stratification have outlined the use of "marked difference" for making visual determinations of racial categories [66, 93, 133], rather than traceable ancestral history. In other words, we suppose these annotations are linked to notions of visuality of making racial determinations, rather than determinations of ethnic origin. While critical race theory is primarily rooted in Western discourse, born of the desire to dismantle racism and white supremacy in the United States [23], we believe it offers a lens for critiquing the structural issues of race in globally-sourced facial analysis databases. Thus, in this paper, we chose to embrace the focus of "race," as defined by critical race scholarship, in the analysis and discussion of image databases. With this in mind, we use the terms "race" and "racial."

4.2.2 *Defining Gender*

Database documentation also used both "gender" and "sex" to refer to the singular concept of an externalized gendered appearance. In this paper, we chose to use the term "gender" to refer to our analysis and discussion of gendered presentation in image databases, aligning ourselves both with trans scholars and concerns that the sex/gender dichotomy is often used to disavow trans identities [29, 47, 151].

4.3 Data Collection and Cleaning

We identified relevant facial analysis databases by examining which databases are being used in recent academic papers. We chose to identify databases used in academic papers because we could then assume they are viewed as useful databases in facial analysis research. To do this, we first created a corpus of 277 research papers that have studied facial analysis. To create this corpus, we started by identifying relevant papers in both the Association for Computing Machinery Digital Library (ACM DL) and Institute of Electrical and Electronics Engineers (IEEE). We chose to use papers published in the ACM and IEEE as they are both two of the largest associations of computing research, and thus contain a great deal of technical research on facial analysis. We scraped 18,661 manuscripts from the ACM DL for papers using the keyword "facial recognition" using Selenium WebDriver [126] (12/12/2019). We also downloaded 4,000 manuscripts from IEEE's Xplore library using the using search terms "facial recognition" and "facial classification" using IEEE's export functionality (12/12/2019)[1]. We aggregated the two datasets from the ACM and IEEE, removing all duplicates, resulting in 16,505 unique manuscripts.

Next, we filtered the manuscripts by "author keywords" for "facial recognition," "face recognition," and "face classification" to ensure the papers were directly relevant to facial analysis research. This resulted in 781 manuscripts. We decided to narrow our corpus to papers published within the last five years, between 2014 and 2019, to ensure both the manageability of the corpus and the modern relevance of the research. This left us with 277 manuscripts from which to manually identify the use of facial analysis databases.

4.4 Identifying Image Databases

We manually coded the remaining 277 manuscripts for referenced databases. We coded any mention of databases, including database creation, use for training, use for evaluation, and databases referenced in literature reviews. Through this process, we identified 160 different databases. We then manually coded each of these 160 databases for what types of media were included in each. This process revealed 15 different types of databases: Image; Video; 3D Model; Image and Video;

[1]The "export" feature can be found on the upper right hand side of the search result page of https://ieeexplore.ieee.org/xplore/

Image and Sketch; Image, Audio, and Video; Video and 3D Model; Image (Eyes); Electrocardiogram (ECG); Audio; Image and Eyes; Image and Audio; Video and Audio; Image and 3D Model; and Image, Sketch, and 3D Model.

Image databases were the most common media type, referenced about five times as often than the next most popular media type, video. Accordingly, we chose to focus on image databases in our analysis of race and gender in databases.

As we analyzed each database, we also eliminated those that did not contain human subjects (e.g., Common Objects in Context (COCO) [97]). We also eliminated two databases for which we could not find explanatory documentation: 5NJJ-YN63 and PCSO_LS Mugshot. Finally, we chose to combine multiple versions of databases together for analysis, treating them as one database (e.g., Yale and Extended Yale B [16]; faces94, faces95, faces96, and grimace [1]) unless there were significant differences between versions (e.g., the introduction of new race and gender categories). This resulted in a final corpus of 92 image databases. For each entry in our corpus, we analyzed original sources such as research papers, websites, and additional supporting documentation.

4.5 Codebook for Analysis

We sought to understand *what purpose* race and gender were meant to serve for facial analysis systems; *which*, if any, sources race and gender were built on; and *what* processes were used to annotate race and gender information. In an initial review of our database sources, we noticed the presence of this information was sporadic, at best. Thus, we developed a simple codebook that was iteratively updated in phases to capture four aspects of the databases:

(1) Whether the database included information about race and/or gender
(2) If race/gender was defined
(3) Whether race/gender was either (1) only provided in the form a summary for the entire database or (2) race and/or gender was annotated at the level of individual images
(4) How race/gender was annotated (when annotations were provided)

We developed this codebook to quantify trends across the databases and to focus our qualitative investigation. We coded the databases using the available documentation about them. Documentation included original research publications, auxiliary materials, websites, posted slide decks, and the databases themselves. The first author developed a codebook by first open-coding the types of identities present in the corpus of databases [49]. Through regular discussion between the first author and the fourth author, the first author went back to develop tighter codes for the how race and gender showed up in the databases (what we refer to as "database types") and how those identities were explained (what we refer to as "sources" and "annotation processes"). After finalizing this codebook, the first, second, and third authors then coded all database materials using the codebook. All coders regularly discussed their thought processes when coding, at the end of which the first author verified the coding of each codebook entry.

We breakdown the codebook in Table 1 in the following sections. For clarity, we provide a number of examples of coded documentation to highlight both the diversity of ways that information was represented across these databases, and how that information was coded.

4.5.1 Race and Gender

For each database, we coded the presence and absence of race and gender. We coded "present" if the database contained race and/or gender. We coded "absent" if it did not.

Race: We coded race as any mention of racial categories, ethnicity, or skin color. The following example represents a snippet we would code as containing race:

Codebook			
Concept	***Code***	***Description***	***Example***
Attributes Present	Race	Race, ethnicity, or skin color	Asian, white, dark skin
	Gender	Gender or sex	Man, female
Database Type	Implicit	The database includes race and/or gender information in the form of demographic distributions	56% female and 44% male
	Explicit	Every image in the database is annotated with race and/or gender information	Images of men are marked with an 'M'
Source	Present	An explanation for how race/gender was defined or derived	A formal citation defining the selected race classification
	Absent	No sources were used to explain the definition of race/gender	
Annotation Process	Present	An explanation for how explicit annotations were conducted	A description of how race categories were annotated by crowdworkers
	Absent	No explanation describing the annotation process	

Table 1. A table showing our codebook. Every database was marked as either Implicit, Explicit, or Neither. All databases were coded with either a 0 (absent) or 1 (present) for both gender and race.

> *We also manually annotated the basic attributes (gender, age (5 ranges) and race) of all RAF faces. ... For racial distribution, there are 77% Caucasian, 8% African-American, and 15% Asian.* —Real-World Affective Faces Database (RAF-DB) [94]

Gender: We coded gender as any mention of gender or sex categories (e.g., gender, sex, men, male, etc.). We also coded proxies of gender, such as familial relationships like mother and daughter. For example, AR Database was coded as including gender based on the following text:

> *Men's image names start with an 'M' symbol and women's images start with an 'W'.* —AR Database [105]

Some databases discussed race and gender, but only in so far as to explicitly state race/gender were not accounted for. For example:

> *Some questions were raised about the age, racial, and sexual distribution of the database. However, at this stage of the program, the key issue was algorithm performance on a database of a large number of individuals.* —FERET [75]

In these instances, we did not code the database as including race and/or gender.

4.5.2 *Implicit and Explicit*

We identified two ways human race and gender are included in databases. Some databases explicitly annotate every image with race and/or gender information. Others, however, only provide race and/or gender information in high-level descriptive statistics for the dataset as a whole. We coded these databases as "explicit" and "implicit" respectively.

4.5.3 *Source for Definition of Race/Gender*

After identifying which databases included race and gender, we coded each for whether the database provided a source for how race and/or gender were defined. We accepted both formal citations and claimed reflexive expertise. We required the source justify the *definitions* of race and gender categories. For example:

> *As prior work has pointed out, skin color alone is not a strong predictor of race, and other features such as facial proportions are important [54, 78, 130, 131]. Face morphology is also relevant for attributes such as age and gender [135]. We incorporated multiple facial coding schemes aimed at capturing facial morphology using craniofacial features [45, 46, 135].* —IBM Diversity in Faces (DiF) [109][2]

4.5.4 *Annotation Practices*

Finally, we coded each explicit database for explanations of how the authors conducted annotations. Our criteria was some form of explanation to how race and/or gender was explicitly annotated. For example, if the authors explained that they visually evaluated each image to make a determination about a race classification:

> *Demographics for the 10k US Adult Faces Database were determined by an Amazon Mechanical Turk demographics study involving 12 workers per face. Amazon Mechanical Turk worker demographics were assembled from demographics surveys attached to the main tasks of Experiments 1 and 2.* —10k US Adult Faces [15]

Based on an analysis of the coded databases, we present our findings in three sections: (1) the purpose and intended use of databases (see section 5); (2) the implicit race and gender features found in databases (see section 6); and (3) the explicit annotations of race and gender found in databases (see section 7).

4.6 Public Availability of Our Dataset

Given that the image databases are publicly available for academic use, we felt it was ethically responsible to publish the corpus of databases examined in this study for the benefit of other researchers, engineers, and the public. We have created an open access spreadsheet of the 92 databases (and associated versions) examined in this paper. This spreadsheet contains our codebook, including tabs for databases we classified as "implicit" and "explicit." We included the titles of original research papers, links to their Google Scholar entries, and the number of citations at the time of data collection. We also included quotes relevant to how race/gender are defined, sourced, and annotated, when available.

We encourage other researchers to use our dataset for additional research and to add new database entries. The dataset is available for download using the following DOI: 10.5281/zenodo.3735400

[2]Note: Numbered citations in quotes throughout this paper has been altered to map to the correct references in the current paper.

5 Contextualizing Race and Gender by Understanding the Intended Purpose of Databases

Image databases serve as an important resource for facial analysis research. Each time a new database is released, the authors are looking to fulfill some need within this community. We observed numerous justifications for the creation of new databases. For example, many databases were looking to continuously expand the number of individual faces available within a single database. We observed three major categories for which image databases were intended to be used for: (1) individual face recognition and verification; (2) image labeling and classification; and (3) for diverse training and evaluation. In these three categories, race and gender were included—or not included—for different reasons. Understanding these reasons contextualizes why we see race and gender manifest in both implicit and explicit ways.

5.1 Face Recognition and Verification

A great deal of database authors described the utility of their database for individual face recognition and verification tasks—that is, tasks meant to match a single individual to a database of images. Race and gender were often implicit in these recognition databases. It is likely that many database authors did not view explicitly annotated information as relevant to the task of face recognition. Matching a single face to a single identity is often viewed as an individual-level task, not requiring additional labels beyond a unique identifier (e.g., a subject's name, a number ID). However, some databases did include race and gender information—typically, along with other attributes like age and facial expressions. These often described the inclusion of such information for the sake of more expansive, more various data. Motivations for variety, however, are not necessarily the same as improving demographic diversity. Databases which sought to increase variety did not mention diversity as a motivation for identity information.

5.2 Image Labeling and Classification

Some databases were meant to aid with image labeling and classification—the assignment of labels to an image based on a database of images with annotations. Explicit race and gender annotations were common in such databases and were used to train a system to classify those annotated race and gender categories. While race and gender classification literature did not make up the majority of our original corpus, databases like Cohn-Kanade (CK) [77] (which did not have explicit annotations) were still used for identity classification tasks (e.g., [11]), suggesting a gap between the intended and actual use of this database.

5.3 Diverse Training and Evaluation

A number of image databases were created to improve the diversity of available faces for training and evaluation, and thus, ideally, mitigate potential representation biases within facial analysis models (e.g., DiF [109], PPB [24]). Such databases were looking to improve conditions for both face recognition and image labeling tasks, but their explicit contribution to the field is motivated by addressing known biases and underrepresentation, allowing for systems to recognize a wider variety of human faces and identity attributes. We saw that most databases utilized gender as a means to "balance" racial diversity—that is, to ensure there are an comparative number of women of a certain race to the number of men of a certain race. Explicit race and gender annotations may or may not be present in databases created for diversity—some chose to provide implicit demographic distribution information instead. Implicit demographic distributions still described the diversity of people represented in a database, while explicit annotations of that diversity could also allow for improved image classifications.

6 Implicit Features

Approximately 64% (n=59) of the 92 image databases we coded did not contain explicit annotations. About 37% (n=34) contained no information about race or gender whatsoever (see Table 2). Approximately 27% (n=25) databases contained implicit information about race and/or gender in the form of demographic distributions. These implicit databases contained descriptive statistics about the race and/or gender of the people featured in the database, but did not annotate that information for each image in the database. Only 4% (n=1) database with implicit data included source information for where demographic categories came from. The other 96% (n=24) databases did not contain any source information underlying demographic descriptions; we generally assumed that the database authors gathered this information directly from subjects or determined subject race/gender themselves.

6.1 Types of Race and Gender Categories

Demographic descriptions manifested semantically in numerous ways. We observed both "gender" (e.g., CASIA-WEBFACE [174]) and "sex" (e.g., NUAA Photograph Imposter Database [160]), as well as both being used interchangeably to refer to the same concept (e.g., HRT Transgender Face Database [102]). Similarly, we observed "race" (e.g., Sheffield (previously UMIST) [57]) and "ethnicity" (e.g., VGGFace2 [26]), as well as both being used interchangeably (e.g., CMU Pose, Illumination, and Expression (CMU PIE) [155], Compound Facial Expressions of Emotion (CFEE) [39]). Underlying these concepts, we also observed numerous instances of categorical labels. For example, CFEE stated:

> *A total of 230 human subjects (130 females; mean age 23; SD 6) were recruited from the university area, receiving a small monetary reward for participating. Most ethnicities and races were included, and Caucasian, Asian, African American, and Hispanic are represented in the database.* —CFEE [39]

In this instance, "female" is used as a default gender, implying there must be another gender accounted for in the demographic distribution (most likely "male," given binary trends). *"Most ethnicities and races"* also similarly insinuates some races are not accounted for, but the authors believe their subject pool accounts for "most" of them. We also found some troubling descriptions, which relied on otherwise criticized or contentious terminology. One of these was in the documentation of the now unavailable Microsoft Celeb (MS-CELEB-1M) [59], which employed the categories "Caucasian," "Mongoloid," and "Negroid." The authors refer to these terms as encompassing *"all the major races in the world"* [59]. It is possible the use of such terms are tied to historic scientific uses of the term to describe physiological differences between races; however, this was the only facial

Race and Gender Representation in Image Database Corpus				
Explicit Annotations			***Implicit Data***	***Neither***
35.9% (33)			27.2% (25)	36.9% (34)
Race	*Gender*	*Both*		
45.5% (15 of 33)	100% (33 of 33)	45.5% (15 of 33)		

Table 2. The number of databases that contained (1) explicit annotations; (2) implicit demographic information; or (3) neither explicit or implicit race and gender information. Each count is out of 92 total image databases.

analysis database we saw use this term, indicating it is likely uncommon in computer vision literature. Such descriptions imply author determinations tied to cultural notions about race. Further, MS-CELEB-1M did not provide detailed distributions of these categories, stating:

> *The diversity (gender, age, profession, race, nationality) of our celebrity list is guaranteed by the large scale of our dataset.* —MS-CELEB-1M [59]

Other databases also claimed to include different genders and races, but did not describe what terms or categories they used. For example, the authors of NUAA wrote in their publication, *"Note that [the database] contains various appearance changes commonly encountered by a face recognition system (e.g., sex, illumination, with/without glasses)"* [160]. However, they did not describe what "sex" looked like in the database.

6.2 Sources for Race and Gender Categories

The only database to contain source material for implicit demographic information was VGGFace2. The authors describe using Freebase knowledge graph to determine the "attribute information such as ethnicity" for the images of IMDB celebrities in their database [26].

Other databases, which we might expect to contain source material based on the outlined methodology, did not. For example, although the Facial Expression Recognition 2013 (FER-2013) Database described using Google search with keywords, they did not describe the process they undertook to define the keywords [111]. So although they stated that *"keywords were combined with words related to gender, age or ethnicity"* [111], it is impossible to tell how the keywords were determined. Similarly, the HRT Database contained no source material on trans people in their definition of transgender as *"someone who under goes a gender transformation via hormone replacement therapy; that is, a male becomes a female by suppressing natural testosterone production and exogenously increasing estrogen"* [102].

6.3 Choosing Implicit Demographics over Explicit Annotations

As we reported in Section 5, databases meant for facial recognition and verification often do not need explicit race or gender annotations to function for their intended purpose, even when they sought to improve diversity. Even databases, which were built specifically in response to human characteristics, did not contain annotations. For example, the HRT Database, which was created for the purpose of identifying individuals across gender transition, was not annotated with gender information about individuals. The HRT Database was particularly unique in comparison to other databases in its treatment of "gender" and "sex." It was also the only database which discussed transgender identities:

> *Gender transformation occurs by down selecting the natural sex hormone of a person in replacement for its opposite. This is known medically as hormone replacement therapy; however, more broadly this can be described as hormone alteration or medical alteration.* —HRT Database [102]

In the HRT Database, transgender faces are problematized for recognition and verification tasks. Gender presentation is thus described not as an identity, but rather a challenge to facial analysis systems.

Often, individual subjects in the database were documented by race and gender to such a degree that descriptive statistics were possible, yet that documentation was never translated into explicit annotations. The NimStim Set of Facial Expressions Database, which also did not explicitly annotate race, but included demographic information in its documentation, stated:

> *A number of features of the set are advantageous for researchers who study face expression processing. Perhaps the most important is the racial diversity of the actors. Studies often*

> *show that the race or ethnicity of a model impacts face processing both behaviorally and in terms of the underlying neurobiology of face processing. This modulation by race or ethnicity is not identified for all populations and may be driven by experience and bias.* —NimStim [163]

The above snippet outlines the reasoning for why NimStim authors intentionally included individuals of multiple racial categories into their database, despite not annotating those features: to improve accuracy and precision for facial recognition tasks. It is possible racial categories were not explicitly annotated, because the authors did not find that information relevant to recognition. FERET, one of the oldest databases, dated to 1993, was the only database we found to provide reasoning for the lack of annotations. On their website, they wrote:

> *Some questions were raised about the age, racial, and sexual distribution of the database. However, at this stage of the program, the key issue was algorithm performance on a database of a large number of individuals.* —FERET [2]

We also witnessed an interesting example of identity-specific licensing agreements in the Iranian Face Database (IFDB), which prohibited the use of women's images in publication. They stated:

> *Some female's images are also provided in this database. These images will never appear in any document of any form.* —IFDB [121]

IFDB was interesting in this regard, as they displayed some concern over the misuse of women's images by third-party researchers and commercial interests. However, their approach to choosing implicit demographic labels of gender also leave the gender of each image up to interpretation from those same third parties. We return to the concept of identity-specific licensing in the Design Considerations (see section 9).

Some databases would include explicit annotations for one feature but not another. For CMU PIE, which included "sex" but did not include race, they also stated: *"At the time of writing, we have not decided whether or not to include the "race" or "ethnicity" of the subjects in the personal attributes"* [155]. The authors did not detail why they had not decided to include race in their annotations. We rarely found explanations about why demographics were included, or not included, annotated, or not annotated.

We did observe instances of third-party researchers annotating databases which originally did not contain explicitly annotated features. For example, Afifi et al. annotated gender for Labelled Faces in the Wild (LFW) [7]. While we did not include third-party annotations of databases in our official analysis, we return to them in the Design Considerations (see section 9).

7 Explicit Features

As shown in Table 2, approximately 36% (n=33) of the 92 databases we analyzed included explicit annotations—either of race or of gender, or of both. About 45% (n=15) of the 33 databases with explicit features included explicit race annotations, while 100% of the 33 databases with explicit annotations included explicit gender annotations.

Of the databases annotated explicitly with racial features, none of the databases with race annotations contained *only* sources (with no annotation information) for how racial determinations were made; 20% (n=3 of all databases with race annotations) contained explanations for how annotation practices were conducted, but no sources; and 20% (n=3) contained both sources and annotation documentation (see Table 3).

Similarly, no databases contained sources without descriptions of the annotation process for gender (of all databases with gender annotations); 6% (n=2) contained descriptions of the annotation process by which images were labelled with gender; 6% (n=2) contained both a source and a description of the annotation process: IBM DiF and PPB. For both race and gender, the databases

which provided sources and/or descriptions of the annotation processes did so with varying levels of rigor. We further illustrate the observed source material and annotation documentation in our database corpus in the following sections.

7.1 Explicit Race Annotations

7.1.1 Types of Race Categories

Like we found in implicit demographic descriptions, race varied widely across the image databases we analyzed. Once more the concept of race was visually described using numerous concepts, like "race" (e.g., MORPH (I & II) [138], Sheffield) and "ethnicity" (e.g., Annotated Facial Landmarks in the Wild (AFLW) [85], FER-2013) and "skin type" (e.g., PPB) or "skin color" (e.g., IBM DiF). Also like with implicit results, we once more observed numerous instances of categorical labels. For example, Radboud Faces Database (RAFD) contained only two racial categories: "Caucasian" and "Moroccan" [91]. Such categories seem more explicitly tied to origin, than visual characteristics of race; yet, their annotations imply visually determinable information. PUBFIG employed four categories: "White," "Asian," "Black," and "Indian" [87]. PUBFIG's categories seem to be determined by visual racial categories, like white, as well as notions of origin, like Indian. We also found that "other" was sometimes utilized as a category (n=3; MORPH-II, KINFACEW, 10K US Adult Faces).

7.1.2 Sources and Annotation Processes for Race Categories

The authors of the 10K US Adult Faces Database defined their racial categories as "White," "Black," "Hispanic," "East Asian," "South Asian," "Middle Eastern," and "other." They explain that they sourced these categories from "common" Amazon Mechanical Turk demographics found in experiments they conducted with Mechanical Turk workers. They compare the demographics of the workers and their database to the United States Census:

> *Demographics for the 10k US Adult Faces Database were determined by an Amazon Mechanical Turk demographics study involving 12 workers per face. Amazon Mechanical Turk worker demographics were assembled from demographics surveys attached to the main tasks of Experiments 1 and 2 ... The 1990 U.S. Census asks about Hispanic origin as a separate question from race, so there is likely overlap with other races.* —10K US Adult Faces [15]

The U.S. Census categories and the selected 10K US Adult Faces Categories do not perfectly align in the paper, making it difficult for readers to discern what decisions were made in collapsing "East Asian," "South Asian," and "Middle Eastern" into simpler categories. Presumably, these categories were used to pre-define their annotation guidelines, which were also conducted using Mechanical Turk. The authors describe the process for which workers were asked to annotate relevant facial attributes:

> *To collect the facial attributes, we conducted a separate AMT survey similar to [88], where each of the 2222 face photographs was annotated by twelve different workers on 19 demographic and facial attributes of relevance for face memorability and face modification. We collected a variety of attributes including demographics such as gender, race and age, physical attributes such as attractiveness, facial hair and make up, and social attributes such as emotional magnitude and friendliness.* —10K US Adult Faces [82]

The PPB database was the first database to explicitly annotate skin tone as a proxy for race, annotating images with two different skin tones: darker and lighter. They explain their decision to use skin tone due to the instability of racial categories:

> *Since race and ethnic labels are unstable, we decided to use skin type as a more visually precise label to measure dataset diversity. Skin type is one phenotypic attribute that can be*

Sources and Annotation Descriptions in Explicit Databases		
	Race	*Gender*
Source	0%	0%
Annotation Description	20% (3 of 15)	6% (2 of 33)
Both Source & Annotation Description	20% (3 of 15)	6% (2 of 33)
Total # Databases with Source/Annotations Description	40% (6 of 15)	12% (4 of 33)

Table 3. The above table shows a count of sources and annotations descriptions in explicitly annotated databases. Only 2 databases—DiF and PPB—contained both sources and annotation explanations for both race and gender.

> *used to more objectively characterize datasets along with eye and nose shapes. Furthermore, skin type was chosen as a phenotypic factor of interest because default camera settings are calibrated to expose lighter-skinned individuals [145]... By labeling faces with skin type, we can increase our understanding of performance on this important phenotypic attribute.* —PPB [24]

PPB uses the Fitzpatrick skin type scale as both a source of categorizing and a guide for annotation [24]. In their annotation practice, they apply the Fitzpatrick scale to each image in their database. The authors write:

> *For the new parliamentarian benchmark, 3 annotators including the authors provided gender and Fitzpatrick labels. A board-certified surgical dermatologist provided the definitive labels for the Fitzpatrick skin type.* —PPB [24]

IBM DiF was the only database which attempted to conceptualize race using multiple phenotypic categories. They employed both skin color and numerous facial coding schemes to determine racial annotations. They derive diverse racial categories from multiple sources: skin tone from Chardon et al. [28]; craniofacial distance from Farkas et al. [46]; and craniofacial ratios from [98]. Like PPB, these different sources also drove the annotation process:

> *As prior work has pointed out, skin color alone is not a strong predictor of race, and other features such as facial proportions are important [50, 54, 130, 131]. Face morphology is also relevant for attributes such as age and gender [135]. We incorporated multiple facial coding schemes aimed at capturing facial morphology using craniofacial features [45, 46, 135].* —IBM DiF [109]

Not all databases which provided annotation information also provided source information. The photos making up MORPH are taken from public records, but they do not describe where they source their concepts of race from. The authors of the MORPH database described their race annotation process in auxiliary materials. The authors, who conducted the annotation, evaluated images of those determined to have "inconsistent race" on a case-by-case basis:

> *Each of the 33 people with inconsistent race was evaluated on a case by case basis. A final decision was made according to one of the following criteria:*
> - *Simple Majority: All images for a given person were assigned the race that appeared at least 50% of the time.*
> - *Visual Estimation: Each person's images were inspected one at a time. We decided the race only if there was a wide consensus among our team members.*

- *Other: For some people (e.g. [sic] those of mixed race) it was difficult to guess their race from the photos, and there was substantial variation in the original dataset. We set the race of all images to Other.* —MORPH [19]

From this description, we can only assume the MORPH authors made subjective decisions about each "inconsistent race" image based on their own perceptions about what race should look like. While this approach is presumably true for many databases with annotated race characteristics, it is made apparent through MORPH's description of their consensus-driven approach.

7.2 Explicit Gender Annotations

7.2.1 Types of Gender Categories

In all gender annotations, gender was only categorized in two ways: "gender" (e.g., LFW, CMU PIE) and "sex" (e.g., NUAA). Similarly, categorical variables took two variations: "male/female" or "man/woman." The majority of the databases used the same schema as CAS-PEAL: "male" and "female" [51]. AR Database annotated each image with an "M" to indicate men and a "W" to indicate woman [105]. We also determined proxies for gender to be explicit gender labels. For example, FAMILY101 used the labels "son," "daughter," "father," "mother," "wife," and "husband" [44]. These annotations could also easily be used for gender classification tasks, thus we determined them to be gender annotations.

A particularly interesting example was PUBFIG, which had two gendered annotations: "male" and "attractive woman," of which there was no associated "female" [87]. The absence of annotations for "female" or "attractive man", however, highlights the culturally-situated values around gender that can emerge within an annotation schema (c.f., [150]).

Given previous literature documenting that automated gender classification only exists as a binary [80, 150], we were not surprised to find almost all databases used only two categories that equated to "male" and "female". The only outlier in this regard was RAF-DB, which contained "5% remains unsure" in their description of gender annotations [94]. While "unsure" still insinuates the individuals in the database should fit into the gender binary, it showcases something beyond "male" and "female." Overall, the following section focuses primarily on how gender categories were determined and how the process of labeling them was discussed.

7.2.2 Sources and Annotation Processes for Gender Categories

Two databases described the annotation process they used to apply gender labels. PUBFIG, which used the labels *"man"* and *"attractive woman,"* used Amazon Mechanical Turk to crowdsource gender labels:

> *Each job was submitted to 3 different workers and only labels where all 3 people agreed were used. In this way, we collected over 125,000 confirmed labels over the course of a month...* —PUBFIG [87]

FAMILY101, which uses gendered labels to describe family members (e.g., son, daughter), used a similar process. The authors selected pre-determined celebrity families, then asked the crowdworkers to find specific images of those families [44]. Other databases used web-based information to determine gender labels. For example, Indian Movie Face Database used IMBD to annotate for the gender of different actors and actresses [153]. However, it is notable that none of the above databases provided sources for how they defined gender for the purposes of annotation.

Only two databases included source information justifying their gender annotation categories: PPB and IBM DiF. These databases also happened to be the only two detailing their gender annotation process. PPB, employed binary "male" and "female" labels, explained their choice in [24]. The authors wrote: *"In this work we use the sex labels of 'male' and 'female' to define gender classes since*

the evaluated benchmarks and classification systems use these binary labels" [24]. PPB used existing database literature to source their binary gender category selection. The authors describe using a combination of *"the name of the parliamentarian, gendered title, prefixes such as Mr [sic] or Ms [sic], and the appearance of the photo"* to label gender for each image [24].

IBM DiF happened to be the only database which incorporated gender into a "facial coding schema" based on sourced craniological features. They reference Rothe et al. [146] as driving their coding scheme for gender. They also describe multiple approaches for annotating gender. In one step, they used automated gender estimation as described in [146]. In another step, they employed human labelers using the crowdsourcing platform Figure Eight. All of the coding schemes, including both automated gender classification and human labeling, are available for each image.

Both PPB and DiF explicitly describe the binary gender annotations in their respective databases as a limitation. The authors of IBM DiF wrote:

> *Note also that the gender categorization in Table 3, as in much of the prior work, uses a binary system for gender classification that corresponds to biological sex – male and female. However, different interpretations of gender in practice can include biological gender, psychological gender and social gender roles. As with race and ethnicity, over-simplification of gender by imposing an incomplete system of categorization can result in face recognition technologies that do not work fairly for all of us.* —IBM DiF [109]

DiF tried to mitigate the binary effect by using an average score of 0 to 1 for each image to *"to predict a continuous value score for gender between 0 and 1, and not just report a binary output"* [109]. However, they also used a "male" versus "female" scale for subjective human-labeled annotations.

8 Discussion

8.1 Moments of Identification: A Machine Learning Approach to Human Identity

Human identity characteristics have become increasingly operationalized and scrutinized within machine learning literature. On one hand, there are attempts to classify attributes like ethnicity [60, 100, 103] and gender [136, 141, 147]. On the other, there are growing concerns about how identity is being represented [18, 80, 150], whether it is fair [58, 108, 134], and what the outcomes are when it is not [14, 122, 157]. Underlying these concerns are the massive amounts of data that machine learning models require for their training and evaluation. For facial analysis models, in particular, this data must contain visual information about human faces. Thus, we see database authors primarily relying on *moments of identification* [61], using the visible, external appearances of faces to make determinations about race and gender. Yet, the nature of identity is shaped by both cultural and historical factors; it is sociohistorical. This nature reveals challenges to assumptions that racial and gender categories throughout these databases are objective in the first place. Given the lack of engagement with sociohistorical theory or deeper notions of invisible, internal race and gender identities, we thus observed race and gender categories—that are socially and historically complex—*portrayed as obvious, static, and apolitical.*

Underlying the three purposes of image databases—recognition, classification, and image diversity—is the utility, reliability, and accuracy of the ground truth data provided. We observed very few instances where database authors documented how they determined information about race and gender in facial images. However, we observed extensive documentation by database authors on the mechanics of lighting, image quality, and camera angles. Such mechanics are generally rooted in agreed upon standards. For example, no one is disputing that angles exist on a 360 degree plane. The lack of similar engagement with ground truth race and gender data undermined the very purpose of image databases to be usable, reliable, and accurate.

Yet, considering the sociohistorical nature of these categories, standardization and benchmarking remains challenging, if not impossible. Race on the one hand demonstrates the challenges when there is a lack of agreed upon standards. However, in contrast, gender shows how even when there is an agreed upon standard, it can be problematic for sociohistorical identity attributes. Instead of aiming for objective standards for classifying race and gender, clear documentation of the tradeoffs and decisions would be a more reasonable and effective approach.

When we examined race, we quickly saw the notion of objectivity degrade. We observed inconsistent practices for identifying race categories, which was otherwise portrayed as objective or obvious. When it came to race, we saw disparate schemas for classifying race and ethnicity. We observed notions of visible race markers (e.g., "Black") and notions of origin (e.g., "Moroccan"), as seen in RAFD [91]). Such inconsistency actually revealed the inherently subjective nature of identifying race from images, and therefore the apparent lack of a benchmark standard as seen in measuring camera angles.

Gender was somewhat opposite compared to race. There is a longstanding practice in database construction to adopt a physiological binary perspective of "male" and "female." We did observe a small number of recent and emerging efforts to address documentation of race and gender in image databases—in particular, with PPB [24] and IBM DiF [109]. Both databases justified their decisions about the race and gender categories they chose and provided explanations about how they went about annotating them. They also weighed the benefits and tradeoffs of various approaches, as seen in their discussions on the limitations of binary gender categories. However, their different schemas highlight the lack of a shared standard for (in this case) race. The complexity they detail in arriving at their classification, however, further highlights the situatedness of sociohistorical attributes, and the need for better guidelines around how to produce such datasets, and what contextual constraints (e.g., cultural context, time, etc.) should be considered when using them. Yet, authors of newer databases (i.e., DiF and PPB) have begun to question how gender is typically viewed in machine learning, even if they have not figured out ways to annotate images beyond it. HCI researchers have also criticized its erasure of trans experiences (e.g., [62, 80, 150]). These criticisms highlight a gap between facial analysis databases and the social realities they are attempting to capture. It may not be possible, or desirable, to develop a shared standard.

When the nature of identity attributes are approached as "common sense," this may also lead to their portrayal as something neutral, objective, obvious, or even irrelevant. As the classifications of race and gender in these databases become folded into actual facial analysis systems, potentials for assessing the impact of classification decisions become increasingly opaque. As other scholars have noted, continuing to embed such limited perspectives into technical systems has the potential to reinforce harmful historical practices of exclusion [14, 80, 150] and even fortify pseudoscientific practices of asserting invisible internal characteristics from visible identifications, like physiognomy [8, 166].

While the sociohistorical nature of race and gender make straightforward and universal database construction unreasonable, these limitations highlight the importance of entirely new lines of scholarship in computer vision addressing how race and gender are encoded into our systems. Specifically, we suggest that facial analysis researchers embrace two practices for situating their databases: (1) embracing positionality; and (2) adopting a sociohistorical perspective when making decisions around classifications of race and gender. In the rest of this section, we provide more details specific to race and gender and then concrete examples as to how researchers may adopt these practices in the Design Considerations (see section 9).

8.2 Positionality: Author and Annotator Classifications of Race and Gender

When annotating image databases, identification was often conducted through an analysis of the visible characteristics of a subject—usually by an author or a crowdworker. Of the databases which did not provide this information, we can assume that race and gender categories are also assigned primarily by visually assessing each image or subject in a database. Through the work of Stuart Hall, we see these as moments of *identification*, the situated construction of identity categories, to include or exclude specific groups of people [61]. Importantly, identifications of race and gender is conducted through the lens of the person doing the identifying and is situated with *"specific modalities of power"* [61]. Identification is colored by one's experiences, interpretations, and perspectives. Without a doubt, the differing perspectives of database authors are why we observed such a wide variety of taxonomies in our data—"gender" versus "sex," and "ethnicity" versus "race."

These categories are *"constructed within, not outside, discourse... in specific historical and institutional sites within specific discursive formations and practices"* [61]. We see in current database practices the collapsing of human identity—which consists of both visible external and invisible internal aspects—into solely the visible. For example, visible gender expression is being used as a proxy for internal gender identity. Beyond the general lack of source material and annotation descriptions in databases, we also observed a lack of acknowledgment of external identification as a subjective process, informed by one's own position and perspective in shaping race and gender categories. Without understanding the position of the author or annotator, the collapse of subject identity is made to appear neutral or objective. Statements like *"most ethnicities and races were included"* (CFEE [39]) and *"the diversity ... is guaranteed by the large scale of our dataset"* (MS-CELEB-1M [59]) are written into database documentation as if the comprehensive diversity of race is objectively possible.

Due to vague documentation, it was also not apparent what visible markers were most salient when authors or annotators were making race and gender determinations. The physical embodiment of identity manifests in numerous ways, both physiological—like skin color, face shape, body type—and expression—like clothing, makeup, and mannerisms. Such visible embodiment is crucial to identity, but it is also not always as simple and static as portrayed in the neutral and unexplained language of most database documentation. It is intertwined with social, cultural, and historical aspects of gender and race—the otherwise invisible aspects of identity. Prior scholars have critiqued the collapse of both the visible and the invisible. For example, Scheuerman et al. assessed the binary output of commercial gender classification systems, noting that, in facial analysis software, "presentation equals gender" [150]. We saw the "objective" portrayal of visible identifications fall apart in cases like MORPH, where the authors described the difficulty *"to guess [some subjects'] race from the photos"* [19]. As facial technologies are often deployed beyond the locale they were initially developed in—often even globally—they enact forms of identification that do not necessarily align with the cultural reality of race and gender of other cultures and histories. Further, they do provide any methods for ensuring that they do align with localized cultures and histories.

The practice of embracing positionality and acknowledging one's perspective as a researcher has been a longstanding practice in feminist epistemologies (e.g., [13, 43, 144]). Positioning race and gender classifications within a discipline, a theory, a history, or one's self would increase the transparency and utility of the database itself. The practice is meant to inform others of the context the research was conducted in and instill trust in the researcher's perspective. In detailing the subjective perspective of identifying race and gender features in images, database authors would make their decisions more transparent to potential users of their databases. This would also allow third-parties to better understand how the database might fit their specific use case. We discuss one approach for doing this in Section 9.1.1.

8.3 Sociohistorical Sourcing: Tying Historical Approaches of Race and Gender Identification to Database Documentation

Identification in technical systems has spanned centuries and geographies. Race and gender have largely been defined by *visible markers of difference*. These differences have then been encoded into numerous technical systems of identification. Given that the premise of facial analysis databases is to enact moments of identification based on the visible features of people, it is imperative that we understand how race and gender have been operationalized in technical infrastructures that predate machine learning. After all, categories like "Negroid," "Black," and "African American" hail from historically evolving notions of both the physicality of race and countries of origin [66, 159]; categories like "male" and "female" have largely been derived from historically entrenched notions of biological sex that erase trans realities [80, 150] while producing normative cisgender ones [25]. This is particularly important given the necessity of incorporating race and gender into system design for mitigating bias and ensuring equal representation of marginalized groups [30, 90, 124].

In order to understand the potential misuse of classifications in facial analysis technologies, we review how race and gender identifications have been used to enact discriminatory political actions. For race, we discuss the evolution of the Census in the United States [156] and the sordid practice of physiognomy [86]. For gender, we observe an ongoing battle with gender classification schemas in the trans rights movement [89, 118]. We review these examples, connecting them with what we observed in the identification practices in image databases. Our goal in drawing these connections is to ensure facial analysis research avoids replicating problematic practices, and mitigates the creation of technical systems that repeat unjust histories.

8.3.1 Histories of Race Identification Embedded into Databases

Some databases in our findings utilized U.S. Census categories for their definitions of race; however, none engaged with how Census categories have been used to count and erase certain people from political participation. This history is crucial to the evolution of the Census categories in the first place. The initial classification of race only had two distinctions—free or slave—which then evolved into the first census groups: European, African, and Native American [132]. Even now, the Census is constantly evolving alongside shifting political agendas and social change. Who gets recognized on the Census determines who is literally counted. Alongside Census-informed categories, we also observed classifications for "Other," which otherwise erase the racial identities of non-classifiable subjects.

Embedding terms like "Negroid" and "Mongoloid" into database documentation, which have associations with histories of scientific racism based on such visible differences [114, 159], insinuates a lack of understanding of categorical oppression. We observed some databases attempting to move away from politicized racial categories, as found in the census, to more static notions of racial affinity: visible skin tone and facial morphology. This form of identification potentially allows for more accurate categorizations of people than subjective racial categories. However, there are tensions between more accurate measurements and the historical practice of physiognomy: the procedure of asserting one's internal character from visible racial and ethnic characteristics. For example, both the segregation in Apartheid South Africa [22] and the extermination of Tutsi people in the Rwandan genocide were based on racial difference through the selected codification of cranial features [86]. Such tensions—between attempts to more objectively increase image diversity and histories of scientific racism—are problematic to actually address. Explanation for choosing to rely on visible difference would help make author intent visible to third-parties.

Beyond critically rethinking when and how to incorporate race categories into databases, we further encourage critical question as to how those categories may be used in working facial

analysis technologies. As we discussed in the related work, annotations of race and gender can aid the classification of minorities, making it easier to track them using facial analysis.

While we encourage thoughtful inclusion of racial diversity into databases, we caution database authors to consider potential physiognomic ties and oppressive outcomes that might result from operationalizing race; we discuss opportunities for mitigating such uses in the Design Considerations (9).

8.3.2 Histories of Gender Identification Embedded into Databases

Like past researchers have discussed of facial analysis outputs [150], the identification of gender in image databases can reify notions of gender as binary, visible, and obvious. We also observed notions of gender tied to archaic notions of women's appearances. Categories like "attractive woman" (PUBFIG) places additional weight on the visibility of images of women subjects, emphasizing beauty standards that are more often applied to women than men [129]. Such perspectives sit in direct opposition of trans activists, as well as feminist scholarship that seeks to imagine gender as an internal, social, and cultural phenomenon [25]. Trans rights movements have been built on fighting restrictive gender markers on government documentation. Mismatches between identity documents can restrict movement between countries [32, 73]; result in differential healthcare access, particularly due to gendered insurance restrictions [81]; and may result in concerning and risky encounters with police and other officials [116, 120]. Yet legal documentation has changed in response to trans movements. For example, non-binary options are becoming increasingly available in certain U.S. states (e.g., Colorado driver's licenses [42]).

Much like with race, such shifts also showcase the fluid and political nature of gender identification. Moreover, they show the mismatch between the standards being employed for gender in databases with the changing standards of gender in other technical systems. The choices authors embrace when making gender identifications necessitate a thoughtful questioning of the role a database will play when incorporated into working facial analysis systems. Like race, gender should be handled with care when operationalized into databases; in particular, we encourage more nuanced ways of representing gender that neither exclude trans identities nor put them at risk.

Given the replication of historical race and gender categories in databases, and their generalizability for uses in unforeseeable large-scale systems, researchers must critically imagine potential misuses. Such imagination is necessary; research has shown that the omission of explicit consideration of race and gender results in disparities (e.g., [30, 90, 124]), and thus, race is still a necessary construct to consider when building databases for machine learning systems. Examining sociohistorical identifications gives researchers the tools to do just that. Given the historical mistreatment of race and gender reviewed in this section, we encourage computer vision researchers to begin incorporating historical perspectives into their documentation. The status quo of database construction and annotation disservices computer vision research. We seek to envision new, human-centered lines of scholarship aimed at capturing the invisible, internal aspects of identity.

9 Design Considerations

We found that there is a general lack of documentation for how race and gender categories are designed and annotated in training and evaluation databases for facial analysis technologies. This finding exposes numerous opportunities. Specifically, there are abundant opportunities to address the two main issues in the previous sections: to provide clear documentation that addresses both (1) the positionality of the authors and annotations and (2) the sociohistorical context of race and gender categories. We then provide three additional design considerations: (1) revisit, revise, or retract existing image databases; (2) creatively incorporate "invisible" aspects of identity; and (3)

explicitly define identity-specific limitations of use. In this section, we detail these five interventions towards promoting growth in the field of computer vision, and machine learning broadly.

9.1 Provide clear and transparent documentation of race and gender that includes positionality and sociohistorical context

We urge database authors to provide more rigorous documentation about their database creation processes. First, they should expound on the decisions they make to include and exclude certain races and genders in their databases. Second, they should describe how they are defining race and gender. Third, they should write rich descriptions of how they annotate race and gender information. For example, whether they annotated images based on based on participant self-identification or based on appearance. They should also provide any guidelines they follow for conducting annotations. For example, what features made an annotator label an image with "woman." Providing documentation on these decision-making processes would make databases more transparent, and thus more usable to third-parties.

9.1.1 Embrace positionality

As race and gender are sociohistorically situated, so too are the perceptions authors and annotators introduce into databases. Including the perspectives, training, and identities authors and annotators bring to image databases would increase the level of transparency currently absent in decision-making processes. Knowing the demographic distribution of authors and annotators is just as useful as knowing the demographic distribution of subjects in the database. There is detailed precedent in other fields for including positionality statements in research. Sociology, anthropology, and increasingly HCI introduce positionality (or reflexivity) into research to ground the choices researchers make when defining research questions, methods, and findings (e.g., [13, 37, 140]. Database authors can include small statements on positionality in their work. Beyond positionality statements, database authors might also consider weaving in smaller nods of positionality into descriptions of the methods used to define and annotate classifications. To accomplish this, database authors might explain their relevant expertise to the task of identity classification. They might also explicitly ask annotators to describe their own race and gender identities, for the sake of understanding how that might shape their annotations. Ogbonnaya-Ogburu, Smith, To, and Toyama also suggest that in addition to these endeavours, researchers ought to be "other-conscious" [125]; in other words, to consider how their work will be viewed by people in other groups than their own, particularly racial minority groups.

It is important to note we are not advocating for the compulsory disclosure of sensitive experiences or marginalized identities. We acknowledge the increased burden of researchers from marginalized identities in self-disclosure, which may risk their personal and professional lives [92]. There is also the valid concern that work conducted by marginalized individuals on topics of identity will be viewed as less scientific and less valid [68, 152]. Rather, we are advocating for increased context setting around the decisions researchers make when constructing and documenting databases, and a deeper attention to documenting the identities of annotators as seen in more research practice (e.g., the reporting of participant demographics).

9.1.2 Incorporate sociohistorical context

Documentation is crucial to understanding the sociohistorical context in which databases are created and annotated. Database authors should explicitly detail the decisions they have made in defining race and gender, as well as the categories they have chosen to represent. As we have demonstrated through our discussion (see section 8.3), not only do cultural definitions of race and gender change over time, but they are linked to sociohistorical modalities of power. How race

and gender concepts are distinguished is contextual to how such categories are being used. Race is often seen as having shared physical traits, whereas ethnicity is seen as shared cultural traits; furthermore, nationality is sometimes is conflated with both [112]. Such classificatory distinctions can also disguise heterogeneity within both "race" and "ethnicity" [56]. Similarly, "gender" and "sex" are often distinguished as two separate concepts. Furthermore, sexologists and activists have criticized the distinction between "sex" and "gender" for erasing intersex bodies [29, 47]. Many trans and gender scholars have since rebutted the gender and sex distinction altogether (e.g., [25, 27]).

When making distinctions, database authors should review theories of race and gender as part of their literature review, and discuss the limitations and tradeoffs of their decisions in their documentation.

9.2 Revisit, revise, or retract existing image databases

During our analysis, we came across databases which third-party researchers had subsequently annotated. For example, although the original LFW database did not have explicit gender annotations, Affi et al. have subsequently annotated this database with gender annotations [7] on their official website. This is a creative way to extend existing databases, which may not contain race and gender information, and improve them. Given the difficulty in accessing the original subjects for most existing databases, it's likely third-parties would still need to adopt methods for external identification of race and gender. In this case, Affi et al. adopted a binary approach to gender in their adaptation of LFW. However, we encourage third-parties to embrace opportunities to introduce alternative, more inclusive annotations. Much like Scheuerman et al. recommended [150], third-party annotators might instead choose to annotate varieties of gender expression (e.g., feminine, masculine; long hair, short hair; makeup, no makeup), rather than perceived gender identity. They might also embed feminist theories into empirical methods. For example, by introducing Standpoint Theory [143] into crowdworking annotations, purposefully recruiting crowdworkers of diverse races and genders to label images, and situating those identities within annotation reports.

We also encourage the research community to revisit and question the utility and inclusivity of previously published databases. For example, we saw that both the MS-Celeb-1M and the HRT Transgender Faces database had been retracted from their websites. They are no longer available for download. It is likely this is due to critical commentary from both researchers (e.g., [80]) and the media (e.g., [74, 115]). Reassessing the validity of databases should be encouraged by the research community, through encouraging both original authors and third parties to publish work evaluating existing databases. The retraction of databases should not be seen as a failure, but instead a contribution to improving computer vision.

9.3 Creatively engage "invisible" aspects of identity

Like other scholars before us (e.g., [80, 150]), we believe that a true representation of race or gender cannot be ascertained without explicitly coupling both the visible—physical embodiment—and the invisible—social and historical realities that shape the internal sense of self. Yet, it is extremely difficult to incorporate the invisible aspects of identity into such a visual medium as computer vision. Those few databases that attempted to fold in the social and internal aspects of gender (PPB and DiF) did so only insofar as to say their databases were limited by being unable to capture this complexity.

Moving forward, we encourage database authors to re-imagine how identity in image databases can also embrace the invisible characteristics of racial and gender identities. Already, there have been attempts to incorporate self-identification into image databases. For example, Scheuerman et al. built an image database using self-annotated gender [150]. This work attempts to encapsulate the invisible in visible images using self-identification of the subjects. Database authors might also

consider methodologies for collecting self-identification directly from subjects—in studio or by survey. We might also consider looking beyond prevailing one-sided moments of identification—where facial analysis systems classify or verify an individual without consent—to interactive systems which allow individuals agency over identification. Such techniques could greatly improve the depth and accuracy of the representations of the identity of the subject.

We acknowledge such approaches are also faced with a large number of limitations. It may be technically infeasible to gather enough images this way; it may also be technically infeasible to use self-annotated annotations, which may be too complex or varied for computer vision systems. As demonstrated by other researchers, the ethical critiques of facial analysis technologies highlighted in this work cues can be extended to other "invisible" aspects of human identity, like personality traits and emotions (e.g., [172])—many of which were also present in the databases we analyzed. Using invisible characteristics, whether that is gender identity or emotion, may still enable problematic applications of facial analysis and must always be critically examined.

9.4 Explicitly define identity-specific limitations of use

We encourage authors to think through the potential implications of their databases being used in facial analysis technologies—in particularly, the misuses. One way of taking caution against oppressive uses of race and gender categories in databases, as demonstrated through our discussion in Section 8.3, is to create licensing agreements that delimit what kinds of uses are acceptable. Many of the databases we examined in this study had licensing agreements for the type of use (e.g., commercial) that were allowed. In some cases, people who want to use these databases must contact the creators before they are given permission to use it.

However, database authors could go further. Specifically, licensing agreements should address acceptable and unacceptable identity-specific uses of databases. We saw this in one database which we examined, the Iranian Face Database (IFDB), which detailed terms of use for the images of women included in the database [121]. While we acknowledge it is not possible to predict all problematic uses of a database, licensing agreements present a first step to protecting both the subjects in the database and the potential targets of facial analysis systems. A significant benefit of licensing agreements is that authors will more easily be able to identify who is using their database and can inform them of changes or even retractions. When it is unclear how much or whether to restrict use of a database, authors should engage with community groups and advocates to determine what uses are appropriate (e.g., [62, 171]).

10 Limitations and Future Work

We acknowledge the limitations of our methods. When database authors did not clearly state how race and gender were derived or how they were annotated, we were left no choice but to read between the lines. Deeper understanding of the motivations and decisions being made by database authors requires an insider perspective; future work would benefit from interviews with database authors, both in academia and in industry contexts. Furthermore, given the propensity of utilizing crowdworking solutions for large scale database annotations, there are immense opportunities for new research on crowdsourced annotation practices. We seek to conduct future work on the positionality of diverse crowdworkers as annotators of race and gender features.

Furthermore, given the sociohistorical power structures that impact groups at the intersection of both race and gender, more work is needed to address theories of intersectionality in facial analysis databases. We observed race and gender being treated as highly disparate identities; they were not addressed as relevant to one another, except when gender was used to balance out racial categories. Future work on more theory-driven approaches to addressing intersectionality [31, 65, 137] in image databases could help alleviate the lack of sociohistorical context we observed in this study.

This work should also consider less U.S. and Westernized views of identity and structural inequality, as the theories used to examine databases in this paper are largely based on Western theories of gender and race.

Additionally, the classification of subjects through computer vision, whether the labels are derived from subjects themselves or otherwise, may simply be viewed as morally objectionable in many circumstances. In particular, when we consider the historical—and contemporary—operationalization of race and gender classification for political means. As such, even in attempts to fold in complex and self-held invisible identities, we must always consider how those databases may be appropriated to accomplish the types oppression we discussed in this study.

11 Conclusion

Emerging research in the realm of FATE (fairness, accountability, transparency, and ethics) has yielded unique insights in improving equity in facial analysis technologies. Specifically, researchers have called for increased engagement with critical scholarship [64] and complex realities of identity [18, 150]. Our study embraces these calls, adopting a critical sociohistorical perspective of race and gender classification to analyze current identity documentation practices in image databases. We examined (1) for what purposes are race and gender included in image databases; (2) what information about race and gender is implicit and what is explicit; (3) what sources are being used to define categories of race and gender; and (4) what annotation practices are being used to identify race and gender in images.

To accomplish this, we analyzed 92 image databases popularly cited in facial analysis literature. We developed a codebook to examine how database authors described their definitions and annotation practices of race and gender. We found that the majority of database authors neither provided sources for which gender and/or race were derived, nor described the annotation practice of identifying race and/or gender in database images. While a small subset of newer image databases are aimed at increasing, in particular, racial diversity and engage more deeply with literature on race and gender identities, they still rely on moments of visible identification that could be used to augment sociohistorical practices of oppression if adopted inappropriately.

We discussed the current state of the art in database construction: apolitical and obvious approaches that erase the subjective reality of external identifications. We highlighted the politicized history of race and gender identifications, including how facial analysis systems are being adapted to expand harmful and oppressive agendas. Throughout this discussion, we highlighted the role of *visible* difference, that otherwise erases the internal experiences of race and gender. We concluded with recommendations for improving approaches to database construction and documentation, including opportunities for authors to mitigate harm to subjects of facial analysis.

12 Acknowledgments

We are grateful for the dedication of the gender and race scholars who have tirelessly worked for decades to uncover the structural inequity embedded in technologies. We thank the reviewers for insightful feedback that helped improve this paper, as well as Brianna Dym and Ellen Simpson who provided feedback on early drafts of this work.

References

[1] [n.d.]. Face Recognition Data. https://cswww.essex.ac.uk/mv/allfaces/

[2] 2017. Face Recognition Technology (FERET). https://www.nist.gov/programs-projects/face-recognition-technology-ferethttps://www.nist.gov/programs-projects/face-recognition-technology-feret{%}0Awww.nist.gov/programs-projects/face-recognition-technology-feretFace

[3] 2017. Unfairness by Algorithm: Distilling the Harms of Automated Decision-Making. (2017).

[4] 2019. Clarifai. https://clarifai.com/

[5] 2019. Facial Recognition on Phones — What Is It and How Does it Work? https://www.xfinity.com/hub/mobile/facial-recognition-on-phone

[6] Salem Hamed Abdurrahim, Salina Abdul Samad, and Aqilah Baseri Huddin. 2018. Review on the effects of age, gender, and race demographics on automatic face recognition. , 1617–1630 pages. https://doi.org/10.1007/s00371-017-1428-z

[7] Mahmoud Afifi and Abdelrahman Abdelhamed. 2019. AFIF4: Deep gender classification based on AdaBoost-based fusion of isolated facial features and foggy faces. *Journal of Visual Communication and Image Representation* 62 (jun 2019), 77–86. https://doi.org/10.1016/j.jvcir.2019.05.001 arXiv:1706.04277

[8] Blaise Agüera y Arcas, Margaret Mitchell, and Alexander Todorov. 2017. Physiognomy's New Clothes. *Medium* (2017). https://medium.com/@blaisea/physiognomys-new-clothes-f2d4b59fdd6a

[9] Linda Martín Alcoff. 2006. *Visible Identities: Race, Gender, and the Self*. Oxford University Press. 1–352 pages. https://doi.org/10.1093/0195137345.001.0001

[10] Thomas C. Anderson. 2000. The body and communities in cyberspace: A Marcellian analysis. *Ethics and Information Technology* 2, 3 (2000), 153–158. https://doi.org/10.1023/A:1010001504963

[11] A. V. Anusha, J. K. JayaSree, Anusree Bhaskar, and R. P. Aneesh. 2017. Facial expression recognition and gender classification using facial patches. In *2016 International Conference on Communication Systems and Networks, ComNet 2016*. Institute of Electrical and Electronics Engineers Inc., 200–204. https://doi.org/10.1109/CSN.2016.7824014

[12] Shwetank Arya, Neeraj Pratap, and Karamjit Bhatia. 2015. Future of Face Recognition: A Review. In *Procedia Computer Science*, Vol. 58. Elsevier, 578–585. https://doi.org/10.1016/j.procs.2015.08.076

[13] Mariam Attia and Julian Edge. 2017. Be(com)ing a reflexive researcher: a developmental approach to research methodology. *Open Review of Educational Research* 4, 1 (jan 2017), 33–45. https://doi.org/10.1080/23265507.2017.1300068

[14] Fabio Bacchini and Ludovica Lorusso. 2019. Race, again: how face recognition technology reinforces racial discrimination. *Journal of Information, Communication and Ethics in Society* 17, 3 (aug 2019), 321–335. https://doi.org/10.1108/JICES-05-2018-0050

[15] Wilma A. Bainbridge, Phillip Isola, and Aude Oliva. 2013. The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General* 142, 4 (2013), 1323–1334. https://doi.org/10.1037/a0033872

[16] Peter N. Belhumeur, Joäo P. Hespanha, and David J. Kriegman. 1997. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19, 7 (1997), 711–720. https://doi.org/10.1109/34.598228

[17] R. K.E. Bellamy, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, Yunfeng Zhang, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, and Sameep Mehta. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development* 63, 4-5 (oct 2019). https://doi.org/10.1147/JRD.2019.2942287 arXiv:1810.01943

[18] Sebastian Benthall and Bruce D. Haynes. 2019. Racial categories in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*. ACM Press, New York, New York, USA, 289–298. https://doi.org/10.1145/3287560.3287575 arXiv:1811.11668

[19] Garrett Bingham and Ben Yip. 2017. *MORPH-II Dataset: Summary and Cleaning*. Technical Report.

[20] Rena Bivens and Oliver L. Haimson. 2016. Baking Gender Into Social Media Design: How Platforms Shape Categories for Users and Advertisers. *Social Media + Society* 2, 4 (nov 2016), 1–12. https://doi.org/10.1177/2056305116672486

[21] Tom Boellstorff. 2008. *Coming of age in second life: An anthropologist explores the virtually human*. Princeton University Press. 1–316 pages. https://doi.org/10.1111/j.1757-6547.2009.00060.x

[22] Geoffrey C. Bowker and Susan Leigh Star. 1999. *Sorting Things Out: Classification and Its Consequences*. MIT Press. https://books.google.com/books/about/Sorting{_}Things{_}Out.html?id=xHlP8WqzizYC

[23] Kevin Brown and Darrell D. Jackson. 2013. The history and conceptual elements of critical race theory. In *Handbook of Critical Race Theory in Education*. Routledge, 9–22. https://doi.org/10.4324/9780203155721

[24] Joy Buolamwini and Timnit Gebru. 2018. *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification* *. Technical Report. 1–15 pages. http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

[25] Judith Butler. 1988. Performative Acts and Gender Constitution: An Essay in Phenomenology and Feminist Theory. *Theatre Journal* 40, 4 (1988), 519. https://doi.org/10.2307/3207893 arXiv:arXiv:1011.1669v3

[26] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. 2018. VGGFace2: A dataset for recognising faces across pose and age. In *Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018*. Institute of Electrical and Electronics Engineers Inc., 67–74. https://doi.org/10.1109/FG.2018.00020 arXiv:1710.08092

[27] Wendy. Cealey Harrison and John. Hood-Williams. 2002. *Beyond sex and gender*. SAGE. 258 pages.

[28] A. Chardon, I. Cretois, and C. Hourseau. 1991. Skin colour typology and suntanning pathways. *International Journal of Cosmetic Science* 13, 4 (1991), 191–208. https://doi.org/10.1111/j.1467-2494.1991.tb00561.x

[29] Cheryl Chase. 1998. Hermaphrodites with attitude: Mapping the emergence of intersex political activism. *GLQ* 4, 2 (oct 1998), 189–211. https://doi.org/10.1215/10642684-4-2-189

[30] Sam Corbett-Davies and Sharad Goel. 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. (jul 2018). https://doi.org/10.1063/1.3627170 arXiv:1808.00023

[31] Kimberle Crenshaw. 1991. Mapping the Margins: Intersectionality, Identity Politics, and Violence Against Women of Color. *Source: Stanford Law Review* 43, 6 (1991), 1241–1299. http://www.jstor.org/stable/1229039

[32] Paisley Currah and Tara Mulqueen. 2011. Securitizing Gender: Identity, Biometrics, and Transgender Bodies at the Airport. *Social Research* 78, 2 (2011), 557–582. https://doi.org/10.1353/sor.2011.0030

[33] Scott Dance. 2019. Maryland set to add 'X' gender designation to driver's licenses under bill by General Assembly. https://www.baltimoresun.com/news/maryland/politics/bs-md-drivers-licenses-20190313-story.html

[34] David Danks and Alex John London. 2017. Algorithmic bias in autonomous systems. In *IJCAI International Joint Conference on Artificial Intelligence*. 4691–4697. https://doi.org/10.24963/ijcai.2017/654

[35] Abhijit Das, Antitza Dantcheva, and Francois Bremond. 2019. Mitigating bias in gender, age and ethnicity classification: A multi-task convolution neural network approach. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 11129 LNCS. Springer, Cham, 573–585. https://doi.org/10.1007/978-3-030-11009-3_35

[36] Elizabeth Dias, Maggie Haberman, and Ellen Almer Durston. 2019. Trump's Order to Combat Anti-Semitism Divides Its Audience: American Jews. https://www.nytimes.com/2019/12/12/us/politics/trump-anti-semitism-jews.html

[37] Lynn Dombrowski, Ellie Harmon, and Sarah Fox. 2016. Social Justice-Oriented Interaction Design: Outlining Key Design Strategies and Commitments. *Proceedings of the 2016 ACM Conference on Designing Interactive Systems* (2016), 656–671. https://doi.org/10.1145/2901790.2901861

[38] Bryan Dosono and Bryan Semaan. 2019. Moderation practices as emotional labor in sustaining online communities: The case of AAPI identity work on reddit. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery. https://doi.org/10.1145/3290605.3300372

[39] Shichuan Du, Yong Tao, and Aleix M. Martinez. 2014. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences of the United States of America* 111, 15 (apr 2014). https://doi.org/10.1073/pnas.1322355111

[40] Brianna Dym, Jed R. Brubaker, Casey Fiesler, and Bryan Semaan. 2019. "Coming Out Okay": Community Narratives for LGBTQ Identity Recovery Work. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (nov 2019), 1–28. https://doi.org/10.1145/3359256

[41] Yael Eisenthal, Gideon Dror, and Eytan Ruppin. 2006. Facial attractiveness: Beauty and the machine. *Neural Computation* 18, 1 (jan 2006), 119–142. https://doi.org/10.1162/089976606774841602

[42] Elise Schmelzer. 2018. Colorado to allow use of X as sex identifier on driver's licenses starting this month.

[43] Kim V.L. England. 1994. Getting personal: Reflexivity, positionality, and feminist research. *Professional Geographer* 46, 1 (1994), 80–89. https://doi.org/10.1111/j.0033-0124.1994.00080.x

[44] Ruogu Fang, Andrew C. Gallagher, Tsuhan Chen, and Alexander Loui. 2013. Kinship classification by modeling facial feature heredity. In *2013 IEEE International Conference on Image Processing, ICIP 2013 - Proceedings*. 2983–2987. https://doi.org/10.1109/ICIP.2013.6738614

[45] Leslie G. Farkas. 1994. *Anthropometry of the head and face.* Raven Press. 405 pages.

[46] Leslie G. Farkas, Marko J. Katic, Christopher R. Forrest, Kurt W. Alt, Ivana Bagič, Georgi Baltadjiev, Eugenia Cunha, Marta Čvičelová, Scott Davies, Ilse Erasmus, Rhonda Gillett-Netting, Karel Hajniš, Arianne Kemkes-Grottenthaler, Irena Khomyakova, Ashizava Kumi, J. Stranger Kgamphe, Nakamura Kayo-Daigo, Thuy Le, Andrzej Malinowski, Marina Negasheva, Sotiris Manolis, Murat Ögetürk, Ramin Parvizrad, Friedrich Rösing, Paresh Sahu, Chiarella Sforza, Stefan Sivkov, Nigar Sultanova, Tatjana Tomazo-Ravnik, Gábor Tóth, Ahmet Uzun, and Eman Yahia. 2005. International anthropometric study of facial morphology in various ethnic groups/races. *Journal of Craniofacial Surgery* 16, 4 (sep 2005), 615–646. https://doi.org/10.1097/01.scs.0000171847.58031.9e

[47] Anne. Fausto-Sterling. 2008. *Sexing the Body: Gender Politics and the Construction of Sexuality.* Basic Books. 487 pages.

[48] Karen E. (Karen Elise) Fields and Barbara Jeanne Fields. 2012. *Racecraft: The Soul of Inequality in American Life.* 302 pages.

[49] Jane Forman and Laura Damschroder. 2007. Qualitative Content Analysis. , 39–62 pages. https://doi.org/10.1016/S1479-3709(07)11003-7

[50] Siyao Fu, Haibo He, and Zeng Guang Hou. 2014. Learning race from face: A survey. , 2483–2509 pages. https://doi.org/10.1109/TPAMI.2014.2321570

[51] Wen Gao, Bo Cao, Shiguang Shan, Xilin Chen, Delong Zhou, Xiaohua Zhang, and Debin Zhao. 2008. The CAS-PEAL large-scale chinese face database and baseline evaluations. *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans* 38, 1 (2008), 149–161. https://doi.org/10.1109/TSMCA.2007.909557

[52] M. Gargesha and S. Panchanathan. 2002. A hybrid technique for facial feature point detection. In *Proceedings of the IEEE Southwest Symposium on Image Analysis and Interpretation*, Vol. 2002-Janua. Institute of Electrical and Electronics Engineers Inc., 134–138. https://doi.org/10.1109/IAI.2002.999905

[53] Shirin Ghaffary and Rani Molla. 2019. Facial recognition: A map of where surveillance technology is in the US. *Vox* (2019). https://www.vox.com/recode/2019/7/18/20698307/facial-recognition-technology-us-government-fight-for-the-future

[54] Alvin G. Goldstein. 1979. Race-related variation of facial features: Anthropometric data I. *Bulletin of the Psychonomic Society* 13, 3 (nov 1979), 187–190. https://doi.org/10.3758/BF03335055

[55] Sixue Gong, Xiaoming Liu, and Anil K. Jain. 2019. DebFace: De-biasing Face Recognition. (nov 2019). arXiv:1911.08080 http://arxiv.org/abs/1911.08080

[56] Lewis R. Gordon. 2007. Thinking through Identities: Black Peoples, Race Labels, and Ethnic Consciousness. In *The Other African Americans: Contemporary African and Caribbean Immigrants in the United States*. Vol. 45. 69–92. https://books.google.com/books?hl=en{&}lr={&}id=RVweAAAAQBAJ{&}oi=fnd{&}pg=PA69{&}dq=lewis+gordon+race{&}ots=z5jl27ODMI{&}sig=D3Nq5Ng{_}orDomTJkFWT9o7yw29Q{#}v=onepage{&}q=lewisgordonrace{&}f=false

[57] Daniel B. Graham and Nigel M. Allinson. 1998. Characterising Virtual Eigensignatures for General Purpose Face Recognition. In *Face Recognition*. Springer Berlin Heidelberg, 446–456. https://doi.org/10.1007/978-3-642-72201-1_25

[58] Patrick Grother, Mei Ngan, and Kayee Hanaoka. [n.d.]. *Face Recognition Vendor Test (FRVT), Part 3: Demographic Effects*. Technical Report. https://doi.org/10.6028/NIST.IR.8280

[59] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. 2016. MS-celeb-1M: A dataset and benchmark for large-scale face recognition. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 9907 LNCS. Springer Verlag, 87–102. https://doi.org/10.1007/978-3-319-46487-9_6 arXiv:1607.08221

[60] Srinivas Gutta, Harry Wechsler, and P. Jonathon Phillips. 1998. Gender and ethnic classification of face images. In *Proceedings - 3rd IEEE International Conference on Automatic Face and Gesture Recognition, FG 1998*. IEEE Comput. Soc, 194–199. https://doi.org/10.1109/AFGR.1998.670948

[61] Stuart Hall. 2012. Introduction: Who Needs 'Identity'? In *Questions of Cultural Identity*. SAGE Publications Ltd, 1–17. https://doi.org/10.4135/9781446221907.n1

[62] Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M Branham. 2018. Gender Recognition or Gender Reductionism? The Social Implications of Automatic Gender Recognition Systems. In *2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*.

[63] Melissa Hamilton. 2019. The sexist algorithm. *Behavioral Sciences and the Law* 37, 2 (mar 2019), 145–157. https://doi.org/10.1002/bsl.2406

[64] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2019. Towards a Critical Race Methodology in Algorithmic Fairness. *FAT** (dec 2019). https://doi.org/10.1145/3351095.3372826 arXiv:1912.03593

[65] Patricia Hill Collins and Sirma Bilge. 2016. *Intersectionality*. 249 pages.

[66] Charles Hirschman. 2004. The origins and demise of the concept of race. *Population and Development Review* 30, 3 (2004), 385–415. https://doi.org/10.1111/j.1728-4457.2004.00021.x

[67] Sean Hollister. 2019. Google contractors reportedly targeted homeless people for Pixel 4 facial recognition. *The Verge* (2019). https://www.theverge.com/2019/10/2/20896181/google-contractor-reportedly-targeted-homeless-people-for-pixel-4-facial-recognition

[68] Kenn Gardner Honeychurch. 1996. Researching dissident subjectivities: Queering the grounds of theory and practice. *Harvard Educational Review* 66, 2 (jun 1996), 339–355. https://doi.org/10.17763/haer.66.2.322km3320m402551

[69] Ayanna Howard and Jason Borenstein. 2018. The Ugly Truth About Ourselves and Our Robot Creations: The Problem of Bias and Social Inequity. *Science and Engineering Ethics* 24, 5 (oct 2018), 1521–1536. https://doi.org/10.1007/s11948-017-9975-2

[70] Xuedong D. Huang, William H. Gates III, Eric J. Horvitz, Joshua T. Goodman, Bradly A. Brunell, Susan T. Dumais, Gary W. Flake, Trenholme J. Griffin, and Oliver Hurst-Hiller. 2006. Targeted advertising in brick-and-mortar establishments. https://patents.google.com/patent/US8725567B2/en

[71] Don Ihde. 2002. *Bodies in technology*. University of Minnesota Press. 155 pages.

[72] Abigail Z. Jacobs and Hanna Wallach. 2019. *Measurement and Fairness*. Technical Report. arXiv:1912.05511 http://arxiv.org/abs/1912.05511

[73] Sandy E. James, Jody L. Herman, Susan Rankin, Mara Keisling, Lisa Mottet, and Ma'ayan Anafi. 2016. *The Report of the 2015 U.S. Transgender Survey*. Technical Report. National Center for Transgender Equality. 298 pages. http://www.transequality.org/sites/default/files/docs/usts/USTSFullReport-FINAL1.6.17.pdf

[74] James Vincent. 2017. Transgender YouTubers had their videos grabbed to train facial recognition software. https://www.theverge.com/2017/8/22/16180080/transgender-youtubers-ai-facial-recognition-dataset

[75] P. Jonathon Phillips, Hyeonjoon Moon, Syed A. Rizvi, and Patrick J. Rauss. 2000. The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22, 10 (2000), 1090–1104. https://doi.org/10.1109/34.879790

[76] Faisal Kamiran and Toon Calders. 2009. Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication, IC4 2009*. https://doi.org/10.1109/IC4.2009.4909197

[77] Takeo Kanade, Jeffrey F Cohn, and Yingli Tian. 2000. Comprehensive database for facial expression analysis. In *Proceedings - 4th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2000*. 46–53. https://doi.org/10.1109/AFGR.2000.840611

[78] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2018. Progressive growing of GANs for improved quality, stability, and variation. In *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR. arXiv:1710.10196

[79] Matthew Kay, Cynthia Matuszek, and Sean A. Munson. 2015. Unequal representation and gender stereotypes in image search results for occupations. In *Conference on Human Factors in Computing Systems - Proceedings*, Vol. 2015-April. ACM Press, New York, New York, USA, 3819–3828. https://doi.org/10.1145/2702123.2702520

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

[81] Liza Khan. 2011. Transgender Health at the Crossroads: Legal Norms, Insurance Markets, and the Threat of Healthcare Reform. *Yale Journal of Health Policy, Law & Ethics* 11, c (2011), 375–418. https://heinonline.org/HOL/Page?handle=hein.journals/yjhple11{&}id=381{&}collection=journals{&}index=

[82] Aditya Khosla, Wilma A Bainbridge, Antonio Torralba, and Aude Oliva. 2013. Modifying the memorability of face photographs. In *Proceedings of the IEEE International Conference on Computer Vision*. 3200–3207. https://doi.org/10.1109/ICCV.2013.397

[83] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A. Efros, and Antonio Torralba. 2012. Undoing the damage of dataset bias. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 7572 LNCS. 158–171. https://doi.org/10.1007/978-3-642-33718-5_12

[84] Brendan F. Klare, Mark J. Burge, Joshua C. Klontz, Richard W. Vorder Bruegge, and Anil K. Jain. 2012. Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security* 7, 6 (2012), 1789–1801. https://doi.org/10.1109/TIFS.2012.2214212

[85] Martin Köstinger, Paul Wohlhart, Peter M. Roth, and Horst Bischof. 2011. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *Proceedings of the IEEE International Conference on Computer Vision*. 2144–2151. https://doi.org/10.1109/ICCVW.2011.6130513

[86] Karen Krüger. 2010. The Destruction of Faces in Rwanda 1994: Mutilation as a Mirror of Racial Ideologies. *L'Europe en Formation* 357, 3 (2010), 91. https://doi.org/10.3917/eufor.357.0091

[87] Neeraj Kumar, Alexander C. Berg, Peter N. Belhumeur, and Shree K. Nayar. 2009. Attribute and simile classifiers for face verification. In *Proceedings of the IEEE International Conference on Computer Vision*. 365–372. https://doi.org/10.1109/ICCV.2009.5459250

[88] Neeraj Kumar, Alexander C Berg, Peter N Belhumeur, and Shree K Nayar. 2009. Attribute and simile classifiers for face verification. In *Proceedings of the IEEE International Conference on Computer Vision*. 365–372. https://doi.org/10.1109/ICCV.2009.5459250

[89] R. Kunzel. 2014. The Flourishing of Transgender Studies. *TSQ: Transgender Studies Quarterly* 1, 1-2 (jan 2014), 285–297. https://doi.org/10.1215/23289252-2399461

[90] Anja Lambrecht and Catherine E. Tucker. 2016. Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads. (2016). https://doi.org/10.2139/ssrn.2852260

[91] Oliver Langner, Ron Dotsch, Gijsbert Bijlstra, Daniel H.J. Wigboldus, Skyler T. Hawk, and Ad van Knippenberg. 2010. Presentation and validation of the radboud faces database. *Cognition and Emotion* 24, 8 (dec 2010), 1377–1388. https://doi.org/10.1080/02699930903485076

[92] Michael C. LaSala, David A. Jenkins, Darrell P. Wheeler, and Karen I. Fredriksen-Goldsen. 2008. LGBT faculty, research, and researchers: Risks and rewards. *Journal of Gay and Lesbian Social Services* 20, 3 (sep 2008), 253–267. https://doi.org/10.1080/10538720802235351

[93] Amanda E Lewis. 2003. Everyday Race-Making: Navigating Racial Boundaries in Schools. , 283–305 pages. https://doi.org/10.1177/0002764203256188

[94] Shan Li and Weihong Deng. 2019. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing* 28, 1 (jan 2019), 356–370. https://doi.org/10.1109/TIP.2018.2868382

[95] James J. Lien, Jeffrey F. Cohn, Takeo Kanade, and Ching Chung Li. 1998. Automated facial expression recognition based on FACS action units. In *Proceedings - 3rd IEEE International Conference on Automatic Face and Gesture Recognition, FG 1998*. IEEE Computer Society, 390–395. https://doi.org/10.1109/AFGR.1998.670980

[96] Annie Lin. 2017. Facial recognition is tracking customers as they shop in stores, tech company says. https://www.cnbc.com/2017/11/23/facial-recognition-is-tracking-customers-as-they-shop-in-stores-tech-company-says.html

[97] Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 8693 LNCS. Springer Verlag, 740–755. https://doi.org/10.1007/978-3-319-10602-1_48 arXiv:1405.0312

[98] Haibin Ling, Stefano Soatto, Narayanan Ramanathan, and David W Jacobs. 2010. Face verification across age progression using discriminative methods. *IEEE Transactions on Information Forensics and Security* 5, 1 (2010), 82–91. https://doi.org/10.1109/TIFS.2009.2038751

[99] John Locke. 1689. Of Identity and Diversity. In *The Works of John Locke, vol. 1 (An Essay concerning Human Understanding Part 1)*.

[100] Xiaoguang Lu and Anil K Jain. 2004. Ethnicity Identification from Face Images. *Proceedings of SPIE* 5404 (2004), 114–123. https://doi.org/10.1117/12.542847

[101] Gayathri Mahalingam and Chandra Kambhamettu. 2011. Can discriminative cues aid face recognition across age?. In *2011 IEEE International Conference on Automatic Face and Gesture Recognition and Workshops, FG 2011*. 206–212. https://doi.org/10.1109/FG.2011.5771399

[102] Gayathri Mahalingam and Karl Ricanek. 2013. Is the eye region more reliable than the face? A preliminary study of face-based recognition on a transgender dataset. In *IEEE 6th International Conference on Biometrics: Theory, Applications and Systems (BTAS 2013)*. IEEE, 1–7. https://doi.org/10.1109/BTAS.2013.6712710

[103] S. Md Mansoor Roomi, S. L. Virasundarii, S. Selvamegala, S. Jeevanandham, and D. Hariharasudhan. 2011. Race classification based on facial features. In *Proceedings - 2011 3rd National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics, NCVPRIPG 2011*. 54–57. https://doi.org/10.1109/NCVPRIPG.2011.19

[104] Avi Marciano. 2019. Reframing biometric surveillance: from a means of inspection to a form of control. *Ethics and Information Technology* 21, 2 (jun 2019), 127–136. https://doi.org/10.1007/s10676-018-9493-1

[105] A.M. Martinez. 1998. The AR face database. *CVC Technical Report 24* (1998). https://doi.org/10.1023/B:VISI.0000029666.37597

[106] Alex Marzano-Lesnevich. 2019. Flying While Trans. https://www.nytimes.com/2019/04/17/opinion/tsa-transgender.html

[107] Matt McFarland. 2016. Terrorist or pedophile? This start-up says it can out secrets by analyzing faces. https://www.washingtonpost.com/news/innovations/wp/2016/05/24/terrorist-or-pedophile-this-start-up-says-it-can-out-secrets-by-analyzing-faces/

[108] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A Survey on Bias and Fairness in Machine Learning. (2019). arXiv:1908.09635 https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencinghttp://arxiv.org/abs/1908.09635

[109] Michele Merler, Nalini Ratha, Rogerio S. Feris, and John R. Smith. 2019. Diversity in Faces. (jan 2019). arXiv:1901.10436 http://arxiv.org/abs/1901.10436

[110] Roberta De Monticelli. 2002. Personal Identity and Depth of the Person: Husserl and the Phenomenological Circles of Munich and Gottingen. In *Phenomenology World-Wide*. Springer Netherlands, 61–74. https://doi.org/10.1007/978-94-007-0473-2_4

[111] A B Moreno and A Sánchez. 2004. GavabDB: A 3D Face Database. *Proc. 2nd COST275 Workshop on Biometrics on the Internet, 2004* (2004), 75–80.

[112] Ann Morning. 2008. Ethnic classification in global perspective: A cross-national survey of the 2000 census round. *Population Research and Policy Review* 27, 2 (2008), 239–272. https://doi.org/10.1007/s11113-007-9062-5

[113] P. Mozur. 2019. One Month, 500,000 Face Scans: How China Is Using A.I. to Profile a Minority. https://www.nytimes.com/2019/04/14/technology/china-surveillance-artificial-intelligence-racial-profiling.html

[114] Carol C. Mukhopadhyay. 2018. Getting Rid of the Word "Caucasian". In *Privilege: A Reader*. Routledge, 231–236. https://doi.org/10.4324/9780429494802-26

[115] Madhumita Murgia. 2019. Who's using your face? The ugly truth about facial recognition |. https://www.ft.com/content/cf19b956-60a2-11e9-b285-3acd5d43599e

[116] Laura Muth. 2018. Why the Gender on My License Is Female Even Though I'm Nonbinary. https://www.allure.com/story/nonbinary-gender-identity-drivers-license

[117] Lisa Nakamura. 2013. *Cybertypes: Race, Ethnicity, and Identity on the Internet*. Taylor and Francis. 1–169 pages. https://doi.org/10.4324/9780203699188

[118] Viviane K. Namaste. 2000. Invisible Lives: The Erasure of Transsexual and Transgendered People. *Contemporary Sociology* 31, 3 (2000), 264. https://doi.org/10.2307/3089651

[119] Srinivas Narayanan. 2019. An Update About Face Recognition on Facebook. https://about.fb.com/news/2019/09/update-face-recognition/

[120] National Center for Transgender Equality. 2015. 2015 US Transgender Survey Report on the Experiences of Black Respondents. (2015), 28 pages. http://www.transequality.org/sites/default/files/docs/usts/USTS-Black-Respondents-Report.pdf

[121] Melika Abbasian Nik, Melika Abbasian Nik, Mohammad Mahdi Dehshibi, and Dr. Azam Bastanfard. 2007. Iranian Face Database and Evaluation with a New Detection Algorithm. (2007). http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.418.771

[122] Safiya Umoja Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism.* 229 pages. https://nyupress.org/books/9781479837243/

[123] Ziad Obermeyer and Sendhil Mullainathan. 2019. Dissecting Racial Bias in an Algorithm that Guides Health Decisions for 70 Million People. In *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19.* ACM Press, New York, New York, USA, 89–89. https://doi.org/10.1145/3287560.3287593

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

[125] Ihudiya Finda Ogbonnaya-ogburu, Angela D R Smith, Alexandra To, and Kentaro Toyama. 2020. Critical Race Theory for HCI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20).* https://doi.org/10.1145/3313831.3376392

[126] OpenQA. 2016. SeleniumHQ - Browser Automation. https://selenium.dev/http://www.seleniumhq.org/

[127] Amanda Phillips. [n.d.]. Making a Face: Quantizing Reality in Character Animation and Customization. In *Gamer Trouble.* 66–99.

[128] Mike Pomranz. 2017. Beer Billboard Uses Facial Recognition to Advertise Only to Women. https://www.foodandwine.com/fwx/drink/beer-billboard-uses-facial-recognition-advertise-only-women

[129] Diane Ponterotto. 2016. Resisting the Male Gaze: Feminist Responses to the "Normatization" of the Female Body in Western Culture. *Journal of International Women's Studies* 17, 1 (2016), 133–151. http://vc.bridgew.edu/jiws

[130] Aurélie Porcheron, Emmanuelle Mauger, Frédérique Soppelsa, Yuli Liu, Liezhong Ge, Olivier Pascalis, Richard Russell, and Frédérique Morizot. 2017. Facial contrast is a cross-cultural cue for perceiving age. *Frontiers in Physiology* 8, JUL (jul 2017), 1208. https://doi.org/10.3389/fpsyg.2017.01208

[131] J. P. Porter and K. L. Olson. 2001. Anthropometric facial analysis of the African American woman. *Archives of facial plastic surgery : official publication for the American Academy of Facial Plastic and Reconstructive Surgery, Inc. and the International Federation of Facial Plastic Surgery Societies* 3, 3 (2001), 191–197. https://doi.org/10.1001/archfaci.3.3.191

[132] Kenneth Prewitt. 2005. Racial classification in America: Where do we go from here? *Daedalus* 134, 1 (2005), 5–17. https://doi.org/10.1162/0011526053124370

[133] Alessandra Raengo. 2013. *On the sleeve of the visual: Race as face value.* 1–232 pages.

[134] Inioluwa Deborah Raji and Joy Buolamwini. 2019. *Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products.* Technical Report. 7 pages. www.aaai.org

[135] Narayanan Ramanathan and Rama Chellappa. 2006. Modeling age progression in young faces. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 1. 387–394. https://doi.org/10.1109/CVPR.2006.187

[136] Arnaud Ramey and Miguel A. Salichs. 2014. Morphological Gender Recognition by a Social Robot and Privacy Concerns. *Proceedings of the 2014 ACM/IEEE International conference on Human-Robot Interaction (HRI '14)* (2014), 272–273. https://doi.org/10.1145/2559636.2563714

[137] Yolanda A. Rankin and Jakita O. Thomas. 2019. Straighten up and fly right: Rethinking intersectionality in HCI research. *Interactions* 26, 6 (2019), 64–68. https://doi.org/10.1145/3363033

[138] Karl Ricanek and Tamirat Tesafaye. 2006. MORPH: A longitudinal image database of normal adult age-progression. In *FGR 2006: Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition*, Vol. 2006. 341–345. https://doi.org/10.1109/FGR.2006.78

[139] Dan Robitzki. 2019. Japanese Taxis Are Using Facial Recognition to Target Ads to Riders. *Futurism* (2019). https://futurism.com/japanese-taxis-facial-recognition-target-ads-riders

[140] Jennifer A. Rode. 2011. Reflexivity in digital anthropology. In *Conference on Human Factors in Computing Systems - Proceedings.* ACM Press, New York, New York, USA, 123–132. https://doi.org/10.1145/1978942.1978961

[141] Pau Rodríguez, Guillem Cucurull, Josep M. Gonfaus, F. Xavier Roca, and Jordi Gonzàlez. 2017. Age and gender recognition in the wild with deep attention. *Pattern Recognition* 72 (dec 2017), 563–571. https://doi.org/10.1016/J.PATCOG.2017.06.028

[142] Yuji Roh, Geon Heo, and Steven Euijong Whang. 2019. A Survey on Data Collection for Machine Learning: A Big Data - AI Integration Perspective. *IEEE Transactions on Knowledge and Data Engineering* (2019), 1–1. https://doi.org/10.1109/tkde.2019.2946162 arXiv:1811.03402

[143] Kristina Rolin. 2009. Standpoint Theory as a Methodology for the Study of Power Relations. *Hypatia* 24, 4 (nov 2009), 218–226. https://doi.org/10.1111/j.1527-2001.2009.01070.x

[144] Gillian Rose. 1997. Situating knowledges: Positionality, reflexivities and other tactics. *Progress in Human Geography* 21, 3 (jun 1997), 305–320. https://doi.org/10.1191/030913297673302122

[145] Lorna Roth. 2009. Looking at Shirley, the Ultimate Norm: Colour Balance, Image Technologies, and Cognitive Equity. *Canadian Journal of Communication* 34, 1 (mar 2009). https://doi.org/10.22230/cjc.2009v34n1a2196

[146] Rasmus Rothe, Radu Timofte, and Luc Van Gool. 2018. Deep Expectation of Real and Apparent Age from a Single Image Without Facial Landmarks. *International Journal of Computer Vision* 126, 2-4 (apr 2018), 144–157. https://doi.org/10.1007/s11263-016-0940-3

[147] Vito Santarcangelo, Giovanni Maria Farinella, and Sebastiano Battiato. 2015. Gender recognition: Methods, datasets and results. In *2015 IEEE International Conference on Multimedia and Expo Workshops, ICMEW 2015*. IEEE, 1–6. https://doi.org/10.1109/ICMEW.2015.7169756

[148] Rachel Savage. 2019. Nonbinary? Intersex? 11 U.S. states issuing third gender IDs. *Reuters* (2019). https://www.reuters.com/article/us-us-lgbt-lawmaking/nonbinary-intersex-11-us-states-issuing-third-gender-ids-idUSKCN1PP2N7https://www.reuters.com/article/us-us-lgbt-lawmaking-idUSKCN1PP2N7

[149] Morgan Klaus Scheuerman, Stacy M Branham, and Foad Hamidi. 2018. Safe Spaces and Safe Places: Unpacking Technology-Mediated Experiences of Safety and Harm with Transgender People. *Proceedings of the ACM on Human-Computer Interaction* 2 (2018), 29.

[150] Morgan Klaus Scheuerman, Jacob M Paul, and Jed R Brubaker. 2019. How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis and Image Labeling Services. 144 (2019), 33. https://doi.org/10.1145/3359246

[151] Julia Serano. 2017. Transgender People and "Biological Sex" Myths. *Medium* (2017). https://medium.com/@juliaserano/transgender-people-and-biological-sex-myths-c2a9bcdb4f4a

[152] Laura Serrant-Green. 2002. Black on black: methodological issues for black researchers working in minority ethnic communities. , 30–44 pages. https://doi.org/10.7748/nr2002.07.9.4.30.c6196

[153] Shankar Setty, Moula Husain, Parisa Beham, Jyothi Gudavalli, Menaka Kandasamy, Radhesyam Vaddi, Vidyagouri Hemadri, J. C. Karure, Raja Raju, B. Rajan, Vijay Kumar, and C. V. Jawahar. 2013. Indian Movie Face Database: A benchmark for face recognition under wide variations. In *2013 4th National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics, NCVPRIPG 2013*. IEEE Computer Society. https://doi.org/10.1109/NCVPRIPG.2013.6776225

[154] Rajeev Sharma, Hankyu Moon, and Namsoon Jung. 2007. Automatic detection and aggregation of demographics and behavior of people. https://patents.google.com/patent/US8351647B2/en

[155] Terence Sim, Simon Baker, and Maan Bsat. 2002. The CMU Pose, Illumination, and Expression (PIE) database. In *Proceedings - 5th IEEE International Conference on Automatic Face Gesture Recognition, FGR 2002*. IEEE Computer Society, 53–58. https://doi.org/10.1109/AFGR.2002.1004130

[156] Mark M. Smith. 2006. *How Race Is Made: Slavery, Segregation, and the Senses*. https://doi.org/10.2307/25094613

[157] Harini Suresh and John V Guttag. 2019. A Framework for Understanding Unintended Consequences of Machine Learning. (2019). arXiv:1901.10002 www.aaai.orghttp://arxiv.org/abs/1901.10002

[158] Zoltán Szlávik and Tamás Szirányi. 2004. Face Analysis Using CNN-UM. In *Proceedings IEEE International Workshop on Cellular Neural Networks and their Applications (CNNA 2004)*. 190–195.

[159] Yasuko Takezawa. 2012. Problems with the Terms : "Caucasoid", "Mongoloid" and "Negroid. *ZINBUN* 43 (2012), 61–68. https://doi.org/10.14989/155688

[160] Xiaoyang Tan, Yi Li, Jun Liu, and Lin Jiang. 2010. Face liveness detection from a single image with sparse low rank bilinear discriminative model. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 6316 LNCS. 504–517. https://doi.org/10.1007/978-3-642-15567-3_37

[161] Tatiana Tommasi, Novi Patricia, Barbara Caputo, and Tinne Tuytelaars. 2017. A deeper look at dataset bias. In *Advances in Computer Vision and Pattern Recognition*. Number 9783319583464. Springer London, 37–55. https://doi.org/10.1007/978-3-319-58347-1_2 arXiv:1505.01257

[162] Antonio Torralba and Alexei A Efros. 2011. Unbiased look at dataset bias. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 1521–1528. https://doi.org/10.1109/CVPR.2011.5995347

[163] Nim Tottenham, James W. Tanaka, Andrew C. Leon, Thomas McCarry, Marcella Nurse, Todd A. Hare, David J. Marcus, Alissa Westerlund, BJ J. Casey, and Charles Nelson. 2009. The NimStim set of facial expressions: Judgments from untrained research participants. *Psychiatry Research* 168, 3 (aug 2009), 242–249. https://doi.org/10.1016/j.psychres.2008.05.006

[164] US Census Bureau. 2019. *2020 Census*. 20 pages. https://www.census.gov/programs-surveys/decennial-census/technical-documentation/questionnaires/2020.htmlhttps://www.census.gov/programs-surveys/decennial-census/2020-census.html

[165] Jennifer Valentino-DeVries. 2020. How the Police Use Facial Recognition, and Where It Falls Short. https://www.nytimes.com/2020/01/12/technology/facial-recognition-police.html

[166] Jeffrey M. Valla, Stephen J. Ceci, and Wendy M. Williams. 2011. The accuracy of inferences about criminality based on facial appearance. *Journal of Social, Evolutionary, and Cultural Psychology* 5, 1 (2011), 66–91. https://doi.org/10.1037/h0099274

[167] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. [n.d.]. *Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations.* Technical Report.

[168] Ylun Wang and Michal Kosinski. 2017. Deep Neural Networks Can Detect Sexual Orientation From Faces. *Journal of personality and social psychology* (2017), 1–47. https://doi.org/10.17605/OSF.IO/HV28A

[169] Rosa Wevers. 2018. Unmasking Biometrics ' Biases: Facing Gender, Race , Class and Ability in Biometric Data Collection. *Tijdschrift voor Mediageschiedenis* 21, 2 (nov 2018), 89–105. https://doi.org/10.18146/tmg21368

[170] Jacob Whitehill and Javier R. Movellan. 2008. Personalized facial attractiveness prediction. In *2008 8th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2008.* https://doi.org/10.1109/AFGR.2008.4813332

[171] Allison Woodruff, Sarah E. Fox, Steven Rousso-Schindler, and Jeff Warshaw. 2018. A qualitative exploration of perceptions of algorithmic fairness. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*, Vol. 2018-April. Association for Computing Machinery, New York, New York, USA, 1–14. https://doi.org/10.1145/3173574.3174230

[172] Niels Wouters, Ryan Kelly, Eduardo Velloso, Katrin Wolf, Hasan Shahid Ferdous, Joshua Newn, Zaher Joukhadar, and Frank Vetere. 2019. Biometric mirror: Exploring values and attitudes towards facial analysis and automated decision-making. In *DIS 2019 - Proceedings of the 2019 ACM Designing Interactive Systems Conference.* Association for Computing Machinery, Inc, 447–461. https://doi.org/10.1145/3322276.3322304

[173] Xiaolin Wu and Xi Zhang. 2016. *Automated Inference on Criminality using Face Images.* Technical Report. arXiv:1611.04135v2

[174] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li. 2014. Learning Face Representation from Scratch. (nov 2014). arXiv:1411.7923 http://arxiv.org/abs/1411.7923

[175] Ning Zhang, Manohar Paluri, Yaniv Taigman, Rob Fergus, and Lubomir Bourdev. 2015. Beyond frontal faces: Improving Person Recognition using multiple cues. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 07-12-June. IEEE Computer Society, 4804–4813. https://doi.org/10.1109/CVPR.2015.7299113 arXiv:1501.05703

Received January 2020; accepted March 2020