Demographic Bias in Biometrics: A Survey on an Emerging Challenge

P. Drozdowski*, C. Rathgeb*, A. Dantcheva†, N. Damert‡, C. Busch*

*da/sec - Biometrics and Internet Security Research Group, Hochschule Darmstadt, Darmstadt, Germany
†Inria, Sophia Antipolis, France
‡Fraunhofer Institute for Computer Graphics Research IGD, Darmstadt, Germany
§Mathematical and Applied Visual Computing, TU Darmstadt, Darmstadt, Germany

Abstract—Systems incorporating biometric technologies have become ubiquitous in personal, commercial, and governmental identity management applications. Both cooperative (e.g. access control) and non-cooperative (e.g. surveillance and forensics) systems have benefited from biometrics. Such systems rely on the uniqueness of certain biological or behavioural characteristics of human beings, which enable for individuals to be reliably recognised using automated algorithms.

Recently, however, there has been a wave of public and academic concerns regarding the existence of systemic bias in automated decision systems (including biometrics). Most prominently, face recognition algorithms have often been labelled as “racist” or “biased” by the media, non-governmental organisations, and researchers alike.

The main contributions of this article are: (1) an overview of the topic of algorithmic bias in the context of biometrics, (2) a comprehensive survey of the existing literature on biometric bias estimation and mitigation, (3) a discussion of the pertinent technical and social matters, and (4) an outline of the remaining challenges and future work items, both from technological and social points of view.

Index Terms—Biometrics, bias, bias estimation, bias mitigation, demographics, fairness.

I. INTRODUCTION

Artificial intelligence systems increasingly support humans in complex decision-making tasks. Domains of interest include learning, problem solving, classifying, as well as making predictions and risk assessments. Automated algorithms have in many cases already outperformed humans and hence are used to support or replace human operators [1]. Those systems, referred to as “automated decision systems”, can yield various benefits, e.g. increased efficiency and decreased monetary costs. At the same time, a number of ethical and legal concerns have been raised, specifically relating to transparency, accountability, explainability, and fairness of such systems [2]. Automated algorithms can be utilised in diverse critical areas such as criminal justice [3], healthcare [4], creditworthiness [5], and others [6], hence often sparking controversial discussions. This article focuses on algorithmic bias and fairness in biometric systems w.r.t. demographic attributes.

In this context, an algorithm is considered to be biased if significant differences in its operation can be observed for different demographic groups of individuals (e.g. females or dark-skinned people), thereby privileging and disadvantaging certain groups of individuals.

A. Motivation

The interest and investment into biometric technologies is large and rapidly growing according to various market value studies [7], [8], [9]. Biometrics are utilised widely by governmental and commercial organisations around the world for purposes such as border control, law enforcement and forensic investigations, voter registration for elections, as well as national identity management systems. Currently, the largest biometric system is operated by the Unique Identification Authority of India, whose national ID system (Aadhaar) accommodates almost the entire Indian population of 1.25 billion enrolled subjects at the time of this writing, see the online dashboard [10] for live data.

In recent years, reports of demographically unfair/biased biometric systems have emerged (see section III), fueling a debate on the use, ethics, and limitations of related technologies between various stakeholders such as the general population, consumer advocates, non-governmental and governmental organisations, academic researchers, and commercial vendors. Such discussions are intense and have even raised demands and considerations that biometric applications should be discontinued in operation, until sufficient privacy protection and demographic bias mitigation can be achieved[11]. Algorithmic bias is considered to be one of the important open challenges in biometrics by Ross et al. [11].

B. Article Contribution and Organisation

In this article, an overview of the emerging challenge of algorithmic bias and fairness in the context of biometric systems is presented. Accordingly, the biometric algorithms which might be susceptible to bias are summarised; furthermore, the existing approaches of bias estimation and bias mitigation are surveyed. The article additionally discusses other pertinent matters, including the potential social impact of bias in biometric systems, as well as the remaining challenges and open issues in this area.

1https://www.banfacialrecognition.com/

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.
The remainder of this article is organised as follows: relevant background information is provided in section II. Section III contains a comprehensive survey of the scientific literature on bias estimation and mitigation in biometric systems. Other relevant matters are discussed in section IV, while concluding remarks and a summary are presented in section V.

II. BACKGROUND

The following subsections provide relevant background information with respect to the topic of bias in automated decision systems in general (subsection II-A) and the basics of biometric systems (subsection II-B). Furthermore, due to the sensitive nature of the matter at hand, subsection II-C outlines the choices made with respect to the nomenclature used throughout the article.

A. Bias in Automated Decision Systems

In recent years, numerous concerns have been raised regarding the accuracy and fairness of automated decision-making systems. For instance, many studies regarding the risk assessment and welfare distribution tools found a number of issues concerning systemic bias and discrimination of the systems’ predictions (e.g., against dark-skinned people). The impact of such automated decisions on the lives of the affected individuals can be tremendous, e.g., being jailed, denied a bail, parole, or welfare payments [2], [3], [12], [13]. Demographics-based bias and discrimination are especially concerning in this context, even if they occur unintentionally. One would intuitively expect that certain decisions be impacted exclusively by hard facts and evidence, and not factors often associated with discrimination – such as sex or race, or other context-specific discriminatory factors. Nonetheless, biases in decision-making are a common occurrence; along with notions of fairness, this topic has been extensively studied from the point of view of various disciplines such as psychology, sociology, statistics, and information theory [14], [15], [16]. Recently, the field of bias and fairness in automated computer algorithms and machine learning has emerged [17], [18].

A good discussion of the topic of bias was provided by Danks and London [19], as well as Friedman and Nis- senbaum [20], both of which explored various sources and types of bias in the context of computer systems. In many cases, bias in the automated decision systems is directly related to the human designers or operators of a system. Semi-automatic decision systems are a good example of this. In such systems, a human decision maker can be aided by an algorithm (e.g., risk-assessment). In such cases, errors in interpretation of the results of the system might occur; in other words, the human might misunderstand or misrepresent the outputs or general functioning principles of an algorithm [21], [22], [23]. Furthermore, it has been shown that humans in general tend to over-estimate the accuracy of their results [24]. While human cognitive biases are an important and actively researched topic, this article focuses exclusively on bias occurring in the context of automated algorithms themselves. Human cognitive biases have been analysed e.g., by Evans [14], whereas bias in human interactions with automated system was explored e.g., by Parasuraman and Manzey [25].

In the context of automated decision algorithms themselves, numerous potential bias causes exist. Most prominently, the training data could be skewed, incomplete, outdated, disproportionate or have embedded historical biases, all of which are detrimental to algorithm training and propagate the biases present in the data. Likewise, the implementation of an algorithm itself could be statistically biased or otherwise flawed in some way, for example due to moral or legal norms, poor design, or data processing steps such as parameter regularisation or smoothing. For more details on the topic of algorithmic bias in general, the reader is referred to e.g., [6], [19], [20]. In the next sections, an introduction to biometric systems is provided, followed by a survey on algorithmic bias in such systems specifically.

B. Biometric Systems

Biometric systems aim at establishing or verifying the identity or demographic attributes of individuals. In the international standard ISO/IEC 2382-37 [26], “biometrics” is defined as: “automated recognition of individuals based on their biological and behavioural characteristics”.

Humans possess, nearly universally, physiological characteristics which are highly distinctive and can therefore be used to distinguish between different individuals with a high degree of confidence. Example images of several prominent biometric characteristics are shown in figure 1.

- (a) Face
- (b) Iris
- (c) Fingerprint
- (d) Veins

Fig. 1: Examples of biometric characteristics (images from publicly available research databases [27], [28], [29], [30]).

Broadly speaking, an automated biometric system consists of: (1) a capture device (e.g., a camera), with which the biometric samples (e.g., images) are acquired; (2) a database which stores the biometric information and other personal data; (3) signal processing algorithms, which estimate the quality of the acquired sample, find the region of interest (e.g., a face), and extract the distinguishing features from it; (4) comparison and decision algorithms, which enable ascertaining of similarity of two biometric samples by comparing the extracted feature vectors and establishing whether or not the two biometric samples belong to the same source.

In the past, biometric systems typically utilised hand-crafted features and algorithms (i.e., texture descriptors, see Liu et al. [31]). Nowadays, the use of machine learning and deep learning has become increasingly popular and successful. Relevant related works include [32], [33], [34], which achieved breakthrough biometric performances in facial recognition.
Furthermore, promising results for deep learning-based fingerprint (see e.g. [35]) and iris (see e.g. [36]) recognition have also been achieved. For a review of deep learning techniques applied within biometrics, the reader is referred to Sundararajan and Woodard [37]. For a highly comprehensive introduction to biometrics, the reader is referred to Li and Jain [38] and the handbook series [39], [40], [41], [42], [43].

C. Nomenclature

In this section, the nomenclature used throughout this article is explained. The authors note that demographic words, groups, and concepts such as “gender”, “sex”, “race”, and “ethnicity” can be extremely divisive and bear a heavy historical, cultural, social, political, or legislative load. The authors do not seek to define or redefine those terms; we merely report on the current state of the research. In the literature surveyed later on in this article, following trends can be distinguished:

1) The terms “gender” and “sex” are often used in a binary and conflated manner. Readers interested in the possible consequences of this narrow approach are referred to [44].

2) Similarly, very often no real distinction between the terms “race” and “ethnicity” is made; moreover, the typical categorisation is very coarse, only allowing for a small and finite (less than ten) possible racial/ethnic categories.

3) In general, and especially in the case of facial biometrics, demographic factors seem to be considered on the phenotypic basis, i.e. concerning the observable traits of the subjects (e.g. colour of the skin or masculine appearance).

Due to the demographic terms carrying a large amount of complexity and potential social divisiveness, the authors do not engage in those debates in this article, and merely reproduce and discuss the technical aspects of the current research. For the sake of consistency, certain decisions regarding the used nomenclature have to be made, especially since the surveyed literature does often seem to use the aforementioned demographic terms ambiguously or interchangeably.

Recently, in the context of biometrics, ISO/IEC has made the following separation [45]

Recently, in the context of biometrics, ISO/IEC has made the following separation [45]: while the term “gender” is defined as “the state of being male or female as it relates to social, cultural or behavioural factors”, the term “sex” is understood as “the state of being male or female as it relates to biological factors such as DNA, anatomy, and physiology”. The report also defines the term “ethnicity” as “the state of belonging to a group with a common origin, set of customs or traditions”, while the term “race” is not defined there. While the cultural and religious norms can certainly affect biometric operations, the surveyed literature mostly considers the appearance-based features and categorisation – hence, the term “race” is used instead of “ethnicity” and the term “sex” is used instead of “gender” in accordance with ISO/IEC 22116 [45]. In the context of biometrics in general, the standardised biometric vocabulary is used, see ISO/IEC 2382-37 [26]. Finally, it is noted that a large part of the surveyed biometric literature follows the notions and metrics regarding evaluation of biometric algorithms irrespective of the chosen biometric characteristic defined in ISO/IEC 19795-1 [46].

Those limitations and imprecisions of the nomenclature notwithstanding, due to the potential of real and disparate impacts [47] of automated decision systems including biometrics, it is imperative to study the bias and fairness of such algorithms w.r.t. the demographic attributes of the population, regardless of their precise definitions.

III. BIAS IN BIOMETRIC SYSTEMS

To facilitate discussions on algorithmic fairness in biometric systems, Howard et al. [48] introduced following two terms:

- **Differential performance** concerns the differences in (genuine and/or impostor) score distributions between the demographic groups. Those effects are closely related to the so-called “biometric menagerie” [49], [50], [51]. While the menagerie describes the score distributions being statistically different for specific individual subjects, the introduced term describes the analogous effect for different demographic groups of subjects.

- **Differential outcomes** relate to the decision results of the biometric system, i.e. the differences in the false-match and false-non-match rates at a specific decision threshold.

Given that these terms have been introduced relatively recently, the vast majority of surveyed literature has not (directly) used them, instead ad hoc methodologies based on existing metrics were used. However, Grother et al. [52] presented a highly comprehensive study of the demographic effects in biometric recognition, conducting their benchmark utilising the terms and notions above. A standardisation effort in this area under the auspices of ISO/IEC is ongoing [45].

Before surveying the literature on bias estimation and mitigation (subsections III-C and III-D, respectively), this section begins with an outline of biometric algorithms which might be affected by bias (subsection III-A), as well as of covariates which might affect them (subsection III-B).

A. Algorithms

Similarly to other automated decision systems, human biases have been shown to exist in the context of biometrics. The so-called “other-race effect” has long been known to affect human ability to recognise faces [53]. As previously stated, the cognitive biases of humans are out of scope for this article, as it focuses on the biases in the algorithms themselves. The processing pipeline of a biometric system can consist of various algorithms depending on the application scenario and the chosen biometric characteristic. Said algorithms might be subject to algorithmic bias w.r.t. certain covariates, which are described in subsection III-B. Below, the most important algorithms used in the context of biometrics are described and visualised conceptually in figure 2.

One of the most prevalent uses of biometrics is recognition. Here, distinguishing features of biometric samples are compared to ascertain their similarity. Such systems typically seek to (1) determine if an individual is who they claim to be (i.e. one-to-one comparison), or (2) to determine the identity of an individual by searching a database (i.e. one-to-many search). Accordingly, the following two scenarios might be used in biometric recognition:
**Verification** Referring to the “process of confirming a biometric claim through biometric comparison” [26], [46].

**Identification** Referring to the “process of searching against a biometric enrolment database to find and return the biometric reference identifier(s) attributable to a single individual” [26], [46].

The biometric samples are a rich source of information beyond the mere identity of the data subject. Another use case of biometrics is the extraction of auxiliary information from a biometric sample, primarily using the following algorithms:

**Classification and estimation** Referring to the process of assigning demographic or other labels to biometric samples [54], [55].

Prior to recognition or classification tasks, the system must acquire and pre-process the biometric sample(s). Here, most prominently, following algorithms might be used:

**Segmentation and feature extraction** Referring to the process of locating the region of interest and extracting a set of biometric features from a biometric sample [38].

**Quality assessment** Referring to the process of quantifying the quality of an acquired biometric sample [56], [57].

**Presentation attack detection (PAD)** Referring to the “automated determination of a presentation attack”, i.e. detecting a “presentation to the biometric data capture subsystem with the goal of interfering with the operation of the biometric system” [58], [59].

B. Covariates

Broadly, three categories of covariates relevant for the effectiveness of the biometric algorithms can be distinguished:

- **Demographic** Referring to e.g. the sex, age, and race of the data subject.
- **Subject-specific** Referring to the behaviour of the subject (e.g. pose or expression, use of accessories such as eyewear or make-up), as well as their interaction with the capture device (e.g. distance from a camera or pressure applied to a touch-based sensor).
- **Environmental** Referring to the effects of the surroundings on the data acquisition process (e.g. illumination, occlusion, resolution of the images captured by the sensor).

Figure 3 shows example images of the aforementioned covariates using the facial biometric characteristic. While there do exist studies that investigate environmental and subject-specific covariates (e.g. [60]), this article concentrates on the demographic covariates.

C. Estimation

Table I summarises the existing research in the area of bias estimation in biometrics. The table is organised conceptually as follows: the studies are divided by biometric characteristic and listed chronologically. The third column lists the algorithms (recall subsection III-A) evaluated by the studies, while the covariates (recall subsection III-B) considered in the studies are listed in the next column. Finally, the last column outlines the key finding(s) of the studies. Wherever possible, those were extracted directly from the abstract or summary sections of the respective studies.

By surveying the existing literature, following trends can be distinguished:
1) Most existing studies conducted the experiments using face-based biometrics. There are significantly fewer studies on other modalities (primarily fingerprint).
2) The majority of studies concentrated on biometric recognition algorithms (primarily verification), followed by quality assessment and classification algorithms.
3) Some scenarios have barely been investigated, e.g. presentation attack detection.
4) The existing studies predominantly considered the sex covariate; the race covariate is also often addressed (possibly due to the recent press coverage [134], [135]). The age covariate is least often considered in the context of bias in the surveyed literature. The impact of ageing on biometric recognition is an active field of research, but out of scope for this article. The interested reader is referred to e.g. [73], [106], [136], [137], [138], [139].
5) Many studies focused on general accuracy rather than distinguishing between false positive and false negative errors. Recent works [48], [52] introduced and used the useful concepts of “false positive differentials” and “false negative differentials” in demographic bias benchmarks.
6) A significant number of studies (e.g. [48], [52], [82]) conducted evaluations on sequestered databases and/or commercial systems. Especially the results of Grother et al. [52] in the context of an evaluation conducted by the National Institute of Standards and Technology (NIST) were valuable due to the realistic/operational nature of the data, the large scale of used databases, as well as the testing of state-of-the-art commercial and academic algorithms. However, reproducing or analysing their results may be impossible due to the unattainability of data and/or tested systems.

Following common findings for the evaluated biometric algorithms can be discerned:

**Recognition** One result which appears to be mostly consistent across surveyed studies is that of worse biometric performance (both in terms of false positives and false negatives) for female subjects (see e.g. [52], [67]). Furthermore, several studies associated race as a major factor influencing biometric performance. However, the results were not attributed to a specific race being inherently more challenging. Rather, the country of software development (and presumably the training data) appears to play a major role; in this context, evidence of the “other-race” effect in facial recognition has been found [65], e.g. algorithms developed in Asia were more easily recognising Asian individuals and conversely algorithms developed in Europe were found to be more easily recognising Caucasians. Finally, the age has been determined to be an important factor as well – especially the very young subjects were a challenge (with effects of ageing also playing a major role). Grother et al. [52] presented hitherto the largest and most comprehensive study of demographic bias in biometric recognition. Their benchmark showed that false-negative differentials usually vary by a factor of less than 3 across the benchmarked algorithms. On the other hand, the false-positive differentials were much more prevalent (albeit not universal) and often larger, *i.e.* varying by two to three orders of magnitude across the benchmarked algorithms\textsuperscript{6}. Most existing studies considered biometric verification, with only a few addressing biometric identification. Estimating bias in biometric identification is non-trivial, due to the contents of the screening database being an additional variable factor susceptible to bias. Specifically, in addition to potential

\textsuperscript{6}Note that this is a very high-level summary to illustrate the general size of the demographic differentials. The experimental results are much more nuanced and complex, as well as dependent on a number of factors in the used data, experimental setup, and the algorithms themselves.

Fig. 3: Example images of covariates which might influence a biometric system utilising facial information (images from a publicly available research database [27]). Black rectangles were added in an effort to respect individual anonymity and privacy.
TABLE I: Summary of studies concerning bias estimation in biometric systems.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Characteristic</th>
<th>Algorithm(s)</th>
<th>Covariates(s)</th>
<th>Key Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beveridge et al. [61]</td>
<td>Face</td>
<td>Verification</td>
<td>Sex, age, race</td>
<td>Better biometric performance for older subjects, males, and East Asians.</td>
</tr>
<tr>
<td>Liu et al. [62]</td>
<td>Face</td>
<td>Verification</td>
<td>Sex, age, race</td>
<td>Meta-analysis of previous studies.</td>
</tr>
<tr>
<td>Guo et al. [63]</td>
<td>Face</td>
<td>Age estimation</td>
<td>Sex, race</td>
<td>Large impact of the training data composition on the system accuracy.</td>
</tr>
<tr>
<td>Grother et al. [64]</td>
<td>Face</td>
<td>Verification</td>
<td>Sex</td>
<td>More false non-matches at fixed FMR for females than for males.</td>
</tr>
<tr>
<td>Phillips et al. [65]</td>
<td>Face</td>
<td>Verification</td>
<td>Race</td>
<td>Varying results depending on the demographic origin of the algorithm and demographic structure of the data subjects.</td>
</tr>
<tr>
<td>O’Toole et al. [66]</td>
<td>Face</td>
<td>Verification</td>
<td>Sex, race</td>
<td>The concept of “yoking” in experimental evaluation to demonstrate the variability of algorithm performance estimates.</td>
</tr>
<tr>
<td>Klare et al. [67]</td>
<td>Face</td>
<td>Verification</td>
<td>Sex, age, race</td>
<td>Lower biometric performance for females, young, and black cohorts.</td>
</tr>
<tr>
<td>Gunes et al. [68]</td>
<td>Face</td>
<td>Verification</td>
<td>Sex, age, race</td>
<td>Better biometric performance for Asian and older subjects.</td>
</tr>
<tr>
<td>Beveridge et al. [69]</td>
<td>Face</td>
<td>Verification</td>
<td>Sex, race</td>
<td>Better biometric performance for males and Asian subjects.</td>
</tr>
<tr>
<td>Ricordeau et al. [70]</td>
<td>Face</td>
<td>Verification</td>
<td>Age</td>
<td>Poor biometric performance for children.</td>
</tr>
<tr>
<td>El Khayat et al. [71]</td>
<td>Face</td>
<td>Verification</td>
<td>Sex, age, race</td>
<td>Lower biometric performance for female, 18-30 age group, and dark-skinned subjects.</td>
</tr>
<tr>
<td>Deb et al. [72]</td>
<td>Face</td>
<td>Verification</td>
<td>Sex, race</td>
<td>Algorithm dependent effects of the covariates.</td>
</tr>
<tr>
<td>BeSicker et al. [73]</td>
<td>Face</td>
<td>Verification</td>
<td>Sex, race</td>
<td>Lower comparison scores for females.</td>
</tr>
<tr>
<td>Buslawnini et al. [74]</td>
<td>Face</td>
<td>Verification</td>
<td>Sex, race, classification</td>
<td>Highest accuracy for males and light-skinned individuals; worst accuracy for dark-skinned females.</td>
</tr>
<tr>
<td>Deb et al. [75]</td>
<td>Face</td>
<td>Verification, identification</td>
<td>Age</td>
<td>Child females easier to recognize than child males.</td>
</tr>
<tr>
<td>Mihalakos et al. [76]</td>
<td>Face</td>
<td>Verification</td>
<td>Age</td>
<td>Large variation of biometric performance across age and ageing factors in children. Poor biometric performance for very young subjects.</td>
</tr>
<tr>
<td>Abbudrham et al. [77]</td>
<td>Face</td>
<td>Verification</td>
<td>Sex, age, race</td>
<td>Lower biometric performance for females, inconsistent results w.r.t. age and race.</td>
</tr>
<tr>
<td>Rhu et al. [78]</td>
<td>Face</td>
<td>Emotion classification</td>
<td>Race</td>
<td>Negative emotions more likely to be assigned to dark-skinned males.</td>
</tr>
<tr>
<td>Lu et al. [79]</td>
<td>Face</td>
<td>Verification</td>
<td>Sex, age, race</td>
<td>Lower biometric performance for females; better biometric performance for middle-aged.</td>
</tr>
<tr>
<td>Raji et al. [80]</td>
<td>Face</td>
<td>Sex and race classification</td>
<td>Sex, race</td>
<td>Lower accuracy for dark-skinned females.</td>
</tr>
<tr>
<td>Sinivas et al. [81]</td>
<td>Face</td>
<td>Verification, identification</td>
<td>Sex, age</td>
<td>Lower biometric performance for females and children.</td>
</tr>
<tr>
<td>Cook et al. [82]</td>
<td>Face</td>
<td>Verification</td>
<td>Sex, age, race</td>
<td>Genuine scores tend to be worse for females than males.</td>
</tr>
<tr>
<td>Hogun et al. [83]</td>
<td>Face</td>
<td>Verification</td>
<td>Sex, race</td>
<td>Highest biometric performance for white males, lowest for Asian females.</td>
</tr>
<tr>
<td>Denton et al. [84]</td>
<td>Face</td>
<td>Classification</td>
<td>CelebA attributes</td>
<td>Generative adversarial model which can reveal biases in a face attribute classifier.</td>
</tr>
<tr>
<td>Garcia et al. [85]</td>
<td>Face</td>
<td>Verification, presentation at-tack detection</td>
<td>Sex, race</td>
<td>Higher inter-subject distance for Caucasians than other groups, morphing attacks more successful for Asian females.</td>
</tr>
<tr>
<td>Nagpal et al. [86]</td>
<td>Face</td>
<td>Verification</td>
<td>Age, race</td>
<td>Training data dependent own-age and own-race effect in deep learning-based systems.</td>
</tr>
<tr>
<td>Kristenpna et al. [87]</td>
<td>Face</td>
<td>Quality, verification</td>
<td>Race</td>
<td>Lower rate of ICA3D compliance [88] for the dark-skinned cohort; fixed decision thresholds not suitable for cross-cohort biometric performance benchmark.</td>
</tr>
<tr>
<td>Muthukumar [89]</td>
<td>Face</td>
<td>Verification</td>
<td>Race</td>
<td>Lower accuracy for dark females; importance of not only skin type, but also illumination in the images on the results.</td>
</tr>
<tr>
<td>Sinivas et al. [90]</td>
<td>Face</td>
<td>Verification, identification</td>
<td>Age</td>
<td>Lower biometric performance for children.</td>
</tr>
<tr>
<td>Vera-Rodriguez et al. [91]</td>
<td>Face</td>
<td>Verification</td>
<td>Sex, race</td>
<td>Lower biometric performance for females.</td>
</tr>
<tr>
<td>Howard et al. [92]</td>
<td>Face</td>
<td>Verification</td>
<td>Sex, age, race</td>
<td>Evaluates effects of population homogeneity on biometric performance.</td>
</tr>
<tr>
<td>Wang et al. [93]</td>
<td>Face</td>
<td>Verification</td>
<td>Race</td>
<td>Higher biometric performance for Caucasians.</td>
</tr>
<tr>
<td>Serna et al. [94]</td>
<td>Face</td>
<td>Verification</td>
<td>Sex, race</td>
<td>Better biometric performance for male Caucasians; large impact of the training data composition on the system accuracy.</td>
</tr>
<tr>
<td>Cuazn et al. [95]</td>
<td>Face</td>
<td>Verification</td>
<td>Sex, race</td>
<td>Higher false match rate for Asians compared to Caucasians at operationally relevant fixed decision thresholds; data-driven anomalies might contribute to system bias.</td>
</tr>
<tr>
<td>Grother et al. [96]</td>
<td>Face</td>
<td>Verification, identification</td>
<td>Sex, age, race</td>
<td>Large-scale benchmark of commercial algorithms. Algorithm dependent false positive differentials w.r.t. race. Consistently elevated false positives for females, elderly and very young subjects. Algorithm specific false negative differentials, also correlated with image quality.</td>
</tr>
<tr>
<td>Robinson et al. [97]</td>
<td>Face</td>
<td>Verification</td>
<td>Sex, race</td>
<td>Highest biometric performance for males and Caucasians.</td>
</tr>
<tr>
<td>Albero et al. [98]</td>
<td>Face</td>
<td>Verification</td>
<td>Sex</td>
<td>Lower biometric performance for females. Negative impact of facial cosmetics on female genuine scores distribution. Minor impact of expression, pose, hair occlusion, and imbalanced datasets on bias.</td>
</tr>
<tr>
<td>Kristenpna et al. [99]</td>
<td>Face</td>
<td>Verification, identification</td>
<td>Race</td>
<td>Lower biometric performance for females; higher false match rate for African-Americans; and higher false non-match rates for Caucasians at fixed, operationally relevant decision threshold.</td>
</tr>
<tr>
<td>Tater et al. [100]</td>
<td>Face</td>
<td>Quality</td>
<td>Age, race</td>
<td>Bias in quality scores for demographic and non-demographic characteristics is significant. Bias is transferred from face recognition to face image quality.</td>
</tr>
<tr>
<td>Hickle et al. [101]</td>
<td>Fingerprint</td>
<td>Quality</td>
<td>Sex</td>
<td>Lower sample quality for females.</td>
</tr>
<tr>
<td>Sicker et al. [102]</td>
<td>Fingerprint</td>
<td>Quality</td>
<td>Age</td>
<td>Lower sample quality for the elderly.</td>
</tr>
<tr>
<td>Modi et al. [103]</td>
<td>Fingerprint</td>
<td>Quality, verification</td>
<td>Age</td>
<td>Lower sample quality and biometric performance for the elderly.</td>
</tr>
<tr>
<td>Modi et al. [104]</td>
<td>Fingerprint</td>
<td>Quality, verification</td>
<td>Age</td>
<td>Lower sample quality and biometric performance for the elderly.</td>
</tr>
<tr>
<td>Frick et al. [105]</td>
<td>Fingerprint</td>
<td>Quality, verification</td>
<td>Sex</td>
<td>Higher sample quality and biometric performance for males.</td>
</tr>
<tr>
<td>O’Connor [106]</td>
<td>Fingerprint</td>
<td>Quality</td>
<td>Sex</td>
<td>Higher sample quality for males; higher biometric performance for females.</td>
</tr>
<tr>
<td>Schumacher et al. [107]</td>
<td>Fingerprint</td>
<td>Quality, verification</td>
<td>Age</td>
<td>Lower sample quality and biometric performance for children.</td>
</tr>
<tr>
<td>Yoon et al. [108]</td>
<td>Fingerprint</td>
<td>Quality, verification</td>
<td>Age, sex, race</td>
<td>Negligible correlations between sample quality and subject age; sex and race have a marginal impact on comparison scores, whereas subject’s age has a non-trivial impact for genuine scores.</td>
</tr>
<tr>
<td>Gathally et al. [109]</td>
<td>Fingerprint</td>
<td>Quality, verification</td>
<td>Sex, age, race</td>
<td>On average, low quality for children under 4 years and elderly (70+ years); medium quality for children between 4 and 12 years. Lowest biometric performance in youngest children, then elderly.</td>
</tr>
<tr>
<td>Peccei et al. [110]</td>
<td>Fingerprint</td>
<td>Quality, verification</td>
<td>Age</td>
<td>Lower sample quality and biometric performance for young children.</td>
</tr>
<tr>
<td>Drzewowski et al. [111]</td>
<td>Fingerprint</td>
<td>Verification</td>
<td>Sex, age</td>
<td>No statistically significant biases detected.</td>
</tr>
<tr>
<td>Fang et al. [112]</td>
<td>Iris</td>
<td>Presentation attack detection</td>
<td>Sex</td>
<td>Better PDR rates for males. Maps differential performance/outcome concepts to PAD.</td>
</tr>
<tr>
<td>Xie et al. [113]</td>
<td>Palmprint</td>
<td>Sex classification</td>
<td>Sex</td>
<td>Higher accuracy for females.</td>
</tr>
<tr>
<td>Ult et al. [114]</td>
<td>Palmprint</td>
<td>Verification</td>
<td>Age</td>
<td>Lower biometric performance for very young subjects.</td>
</tr>
<tr>
<td>Brandalo et al. [115]</td>
<td>Unconstrained</td>
<td>Pedestrian detection</td>
<td>Sex, age</td>
<td>Higher miss rate for children.</td>
</tr>
</tbody>
</table>

Biases in the biometric algorithms themselves, certain biases stemming from data acquisition might occur and be propagated (e.g. historical and societal biases having impact on the demographic composition of a criminal database). Consequently, demographic bias estimation in biometric identification is an interesting and important item for future research.

Classification and estimation Scientific literature predominantly studied face as the biometric characteristic, since the facial region contains rich information from which demographic attributes can be estimated. Several studies showed substantial impact of sex and race on the accuracy of demographic attribute classification. In particular, numerous commercial algorithms exhibited significantly lower accuracy w.r.t. dark-skinned female subjects (see e.g. [74], [80]). Research on classification of sex from iris and periocular images exists, but biases in those algorithms have not yet been studied. Additionally, it is not clear if such classifiers rely on actual anatomical properties or merely the application of mascara [140].

Quality assessment Most existing studies conducted experiments using fingerprint-based biometrics. This could be partially caused by the standardisation of reliable fingerprint quality assessment metrics [141], whereas this remains an open challenge for the face characteristic [142]. The existing fingerprint quality assessment...
In addition, it is necessary to point out potential issues in surveyed studies, such as:

- **Differences in experimental setups**, used toolchains and datasets, training-testing data partitioning, imbalanced datasets etc.
- **Statistical significance** of the results due to relatively small size of the used datasets in most cases (except e.g. [52], [107]).
- **Lack of a single definition of bias/fairness** (see also subsection IV-A), as well as a standardised methodology and metrics for conducting evaluations.

- **Difficulty of sufficiently isolating the influence** of demographic factors from other important covariates (e.g. pose and illumination).
- **Potential for bias propagation** from previous steps of the pipeline (e.g. data acquisition).

Nevertheless, some results appear to be intuitive, e.g. worse accuracies for women. These could be due to numerous reasons, such as: larger intra-class variations due to make-up [147], occlusion by hairstyle and accessories, or pose differences due to women being shorter than men and cameras being calibrated with the height of men. Likewise, lower sample quality of infant fingerprints makes sense due to anatomical constraints and the fact that the size of the fingerprint area is considered as a relevant factor for fingerprint sample quality. In order to acquire high-quality fingerprint samples from very young data subjects, specialised hardware may be necessary (see e.g. [148]).

### D. Mitigation

Table II summarises the existing research in the area of bias mitigation in biometrics. Similarly to above, related work here focuses predominantly on face as biometric characteristic. In this context, mainly recognition and classification algorithms have been analysed. Generally speaking, the existing approaches can be assigned to following categories:

**Training** Learning-based methods have experienced a tremendous growth in accuracy and popularity in recent years. As such, the training step is of critical importance for the used systems and mitigation of demographic bias. The existing techniques mainly rely on demographically balanced training datasets (e.g. [92]) and synthetic data to enhance the training datasets (e.g. [125]), as well as learning specialised loss or similarity functions (e.g. [130]). A number of balanced training datasets has been released to the research community, as shown in table III.
Dynamic selection Deviating from preventing demographic bias, some methods attempted to employ a bias-aware approach. Examples in this category include dynamic selection of the recognition algorithms (e.g. [63]) or decision thresholds (e.g. [87]) based on demographic attributes of the individual subjects.

In addition to the categories above, other approaches may be considered in the context of bias mitigation. For example, modelling of factors such as fingerprint growth can be used to improve the biometric recognition performance for children (see e.g. [109]) and to mitigate the effects of ageing (see e.g. [149]). Other examples include de-identification and anonymisation methods (see e.g. [150], [151]), whose primary use case is privacy-protection in biometrics. Such methods aim to remove, change, or obfuscate certain information (e.g. demographics) either from the image (e.g. [152]) or feature (e.g. [120], [153]) domain, often through a form of adversarial learning. One could hypothesise that a system trained on such data might not exhibit biases w.r.t. to the de-identified demographic covariates. However, the validity of such hypotheses has not yet been ascertained experimentally.

TABLE III: Summary of existing datasets for bias-related research in biometrics.

<table>
<thead>
<tr>
<th>Reference</th>
<th>Characteristic</th>
<th>Size (images)</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bucak et al. [154]</td>
<td>Face</td>
<td>55,134</td>
<td>Ageing research database with demographic labels.</td>
</tr>
<tr>
<td>Azizpour et al. [155]</td>
<td>Face</td>
<td>946</td>
<td>Subset of FERET dataset balanced w.r.t. sex.</td>
</tr>
<tr>
<td>Bashamwini et al. [74]</td>
<td>Face</td>
<td>1,270</td>
<td>Images of parliamentarians balanced w.r.t. sex and age. One image per subject, i.e. not suitable for biometric recognition.</td>
</tr>
<tr>
<td>Alvi et al. [118]</td>
<td>Face</td>
<td>14,000</td>
<td>Scraped images balanced w.r.t. race.</td>
</tr>
<tr>
<td>Alvi et al. [118]</td>
<td>Face</td>
<td>60,000</td>
<td>Subset of IJBDB dataset balanced w.r.t. sex and race.</td>
</tr>
<tr>
<td>Morales et al. [153]</td>
<td>Face</td>
<td>139,677</td>
<td>Subset of MegaFace dataset balanced w.r.t. sex and race.</td>
</tr>
<tr>
<td>Merler et al. [156]</td>
<td>Face</td>
<td>964,873</td>
<td>Demographic and geometric annotations for selected images from YFCC-100M dataset.</td>
</tr>
<tr>
<td>Hupont et al. [83]</td>
<td>Face</td>
<td>10,800</td>
<td>Subset of CFV and VGG datasets balanced w.r.t. sex and race.</td>
</tr>
<tr>
<td>Kärkkäinen et al. [157]</td>
<td>Face</td>
<td>108,501</td>
<td>Subset of YFCC-100M dataset balanced w.r.t. sex, race, and age.</td>
</tr>
<tr>
<td>Wang et al. [92]</td>
<td>Face</td>
<td>40,607</td>
<td>Subset of MSc-Celeb-1M dataset balanced w.r.t. race.</td>
</tr>
<tr>
<td>Robinson et al. [95]</td>
<td>Face</td>
<td>20,000</td>
<td>Subset of IJCB dataset balanced w.r.t. sex and race.</td>
</tr>
<tr>
<td>Albiero et al. [96]</td>
<td>Face</td>
<td>42,134</td>
<td>Subset of AFD dataset balanced w.r.t. sex.</td>
</tr>
</tbody>
</table>

IV. DISCUSSION

In this section, several issues relevant to the topic of this article are discussed. Concretely, subsection IV-A addresses the topic of algorithmic fairness in general, while subsection IV-B does so in the context of biometrics specifically. Subsection IV-C illustrates the importance of further research on algorithmic bias and fairness in biometrics by describing the social impact of demographically biased systems.

A. Algorithmic Fairness in General

The challenge of fairness is common in machine learning and computer vision, i.e. it is by no means limited to biometrics. A survey focusing on issues and challenges associated with algorithmic fairness was conducted among industry practitioners by Holstein et al. [158]. For a comprehensive overview of bias in automated algorithms in general, the reader is referred to e.g. [18], [159]. In addition to algorithmic fairness, algorithmic transparency, explainability, interpretability, and accountability (see e.g. [160], [161], [162], [163]) have also been heavily researched in recent years both from the technical and social perspective. The current research in the area of algorithmic fairness concentrates on the following topics:

- Theoretical and formal definitions of bias and fairness (see e.g. [18], [164], [165]).
- Fairness metrics, software, and benchmarks (see e.g. [166], [167], [168]).
- Societal, ethical, and legal aspects of algorithmic decision-making and fairness therein (see e.g. [1], [169], [170], [171], [172]).
- Estimation and mitigation of bias in algorithms and datasets (see e.g. [173], [174], [175], [176], [177], [178]).

Despite decades of research, there exists no single agreed coherent definition of algorithmic fairness. In fact, dozens of formal definitions (see e.g. [164], [165]) have been proposed to address different situations and possible criteria of fairness7. Certain definitions, which are commonly used and advocated for, are even provably mutually exclusive [179]. Therefore, depending on the definition of fairness one chooses to adopt, a system can effectively always be shown to exhibit some form of bias. As such, the “correct” approach is essentially application-dependent. This in turn necessitates a keen domain knowledge and awareness of those issues from the system operators and stakeholders, as they need to select the definitions and metrics of fairness relevant to their particular use case. Research in this area strongly suggests that the notion of fairness in machine learning is context-sensitive [180], [181]; this presumably also applies to the field of biometrics, especially for machine learning-based systems. In the next subsection, the notions of fairness and bias are discussed in the context of biometrics specifically based on the literature surveyed in section III.

B. Algorithmic Fairness in Biometrics

Although the topic of demographic bias and fairness in biometrics has emerged relatively recently, it has quickly established itself as an important and popular research area. Several high-ranking conferences featured special sessions8,9,10, NIST conducted large-scale evaluations [52], while ISO/IEC is currently preparing a technical report on this subject [45]. Likewise, a significant number of scientific publications has appeared on this topic (surveyed in section III). Existing studies concentrated on face-based biometrics – more research is urgently needed for other biometric characteristics, e.g. fingerprints [182].

7See also https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-30f8ba1040ce and https://fairmlbook.org/ for visual tutorials on bias and fairness in machine learning.
8https://sites.google.com/view/wacv2020demographics
9https://sites.google.com/site/ceccvbeta2018

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.
Existing studies primarily address the following aspects:

1) Evaluations with the aim of quantitatively ascertaining the degree of demographic bias in various biometric algorithms.
2) Methods which seek to mitigate the effects of demographic bias in various biometric algorithms.

Existing bias estimation studies have uncovered new trends w.r.t. algorithmic bias and fairness in biometric algorithms (recall subsection III-C). However, it should be noted, that:

1) In many cases the biases were algorithm-specific, i.e. while given the same benchmark-dataset some algorithms exhibited a bias (e.g. lower biometric performance for a certain demographic group), others did not. In aggregate, however, the existing studies did seem to agree on certain points, as described in subsection III-C.
2) While a high relative increase in error rates for a certain demographic group may appear quite substantial, its importance in absolute terms could be negligible, especially for very accurate algorithms which hardly make any errors whatsoever [52].

Those caveats notwithstanding, the commitment of the academic researchers and commercial vendors to researching algorithmic fairness is especially important for the public perception of biometric technologies. The field of algorithmic fairness in the context of biometrics is in its infancy and a large number of issues are yet to be comprehensively addressed (cf. subsection IV-A):

1) Limited theoretical work has been conducted in this field specifically focusing on biometrics. Indeed, the majority of publications surveyed in section III do not approach the notions of bias and fairness rigorously; rather, they tend to concentrate on an equivalent of some of the simpler statistical definitions, such as group fairness and error rate parity. Extending the existing estimation and mitigation works, for example to consider other and more complex notions of fairness (see e.g. [129]) could be seen as important future work in the field. Likewise, investigating trade-offs between biometric performance, fairness, user experience, social perceptions, monetary costs, and other aspects of the biometric systems might be of interest.
2) In addition to empiric studies (especially in the case of bias mitigation, see subsection III-D), stricter theoretical approaches need to be pursued in order to provably demonstrate the bias-mitigating properties of the proposed methods.
3) Isolating the effects of the demographic factors from other confounding factors (i.e. the environmental and subject-specific covariates, such as illumination and use of accessories) is a challenging task, which is not sufficiently addressed in many existing studies. An example of a study which partially addressed those issues in a systematic manner is the work of Grother et al. [52].
4) More complex analyses based on demographic attributes and combinations thereof (intersectionality) could be conducted for a more detailed and nuanced view of demographic biases in biometric systems.
5) Comprehensive independent benchmarks utilising various algorithmic fairness measurement methodologies and metrics are, as of yet, lacking. Only recently, in [52], first independent benchmarks of biometric recognition algorithms have been conducted. Similar and more extensive benchmarks for other biometric algorithms (recall subsection III-A) are needed.
6) Large-scale datasets designed specifically for bias-related research need to be collected. The existing datasets only pertain to face-based biometrics (see table III).

Humans are known to exhibit a broad range of biases [14], [25]. The influence of those factors on biometric algorithm design, interactions with and use of biometric systems, as well as perceptions of biometric systems could be investigated.

8) Most of the surveyed studies did not explicitly provide information about ethics approval. Future works could improve on those practices, especially considering the sensitive nature of the research topic at hand.

In the next subsection, the possible consequences of failing to appropriately address the issues of algorithmic fairness in biometrics are discussed.

C. Social Impact

Numerous studies described the potential of real harms as a consequence of biased algorithmic decision-making systems [169], [183] in general. Regarding biometric systems in particular, facial recognition technologies have been the main focus of such discussions (see e.g. [184]). Concerning the notions of bias and fairness, in addition to being context-sensitive (recall subsection IV-A), one might argue the impact assessments to also be purpose-sensitive. Specifically, depending on application scenario, the impact and importance of systemic biases might differ significantly. As an example, consider an application of biometrics in cooperative access control systems or personal devices. A demographic bias in such a system might cause a certain demographic group to be inconvenienced through additional authentication attempt(s) being necessary due to false negative errors. On the other hand, the stakes are much higher in e.g. a state surveillance scenario. There, demographic biases could directly cause substantial personal harms, e.g. higher (unjustified) arrest rates [12], due to false positive errors. At the same time, it is also clear that biometric recognition technology can be highly accurate. Taking the recently contested facial recognition as an example, given prerequisites such as a high-resolution camera, proper lighting and image quality controls, as well as high-quality comparison algorithms, the absolute error rates can become vanishingly small [52], thereby potentially rendering the relative imbalance of error rates across demographic groups insignificant.

It should be noted that there are no indications of the algorithmic biases in biometrics being deliberately put into the algorithms by design; rather, they are typically a result of the used training data and other factors. In any case, one should also be mindful, that as any technology, biometrics could be used in malicious or dystopian ways (e.g. privacy violations through mass-surveillance [185] or “crime prediction” [186]). Consequently, a framework for human impact
assessments [187] should be developed for biometrics as soon as possible. A pro-active and cognizant approach could foster awareness among the citizens and policymakers, as well as contribute to minimising potential negative perception of biometric technology and innovation by individuals and society as a whole.

In a broader context, algorithmic bias and fairness is one of the topics in the larger discourse on ethical design in artificial intelligence (AI) systems [188], most prominently encompassing:

- Transparency,
- Accountability,
- Explainability, and
- Fairness.

Currently, the legal and societal scrutiny of the technologies utilising automated decision systems seems to be insufficient. However, recent legislation in the European Union [189], [190] constitutes a step in that direction. Below, several social and technological provisions, which might be considered in this context, are listed.

- Carefully selecting the data used to train the algorithms is the first and perhaps the most important step: inherent biases in training data should be avoided wherever possible. Furthermore, the size of the dataset matters – some systems have been reported to be trained on very small datasets (in the order of thousands of items), which is usually wholly insufficient to show that an approach generalises well.
- Higher degree of transparency and/or independent insight into data and algorithms, as well as validation of the results could be established to foster the public trust and acceptance of the systems.
- Thresholds for acceptable accuracy (i.e. how much the systems can err) could be established legally (potentially in a system purpose-sensitive manner), as well as reviewed and validated periodically.
- Special training of the systems’ personnel could be established to make them aware of the potential issues and to establish proper protocols for dealing with them.
- Due diligence could be legally expected from vendors of such systems, i.e. in reasonably ensuring some or all of aforementioned matters and rectifying problems as they come up. Additionally, certain accountability provisions could be incorporated to further facilitate this.

The issues of fairness (including algorithmic fairness) are complicated from the point of view of the legislation – a somewhat deep understanding of statistics, formal fairness definitions, and other concepts is essential for an informed discourse. Furthermore, the ethical and moral perceptions and decisions are not uniform across different population demographics and by geographical location (see e.g. Awad et al. [191]). This reinforces an important dilemma regarding the regulation of automated decision systems – since many situations are morally and ethically ambiguous to humans, how should they be able to encode ethical decision-making into laws? Once that issue is somehow surmounted, there also remains the issue of feasibility of technical solutions, as described in the previous two subsections.

Currently, many laws and rules exist (international treaties, constitutions of many countries, and employment law) which aim to protect against generic discrimination on the basis of demographics [192]. However, historically, the enforcement of those has been fraught with difficulties and controversies. In this context, the algorithmic decision systems are merely one of the most recent and technologically advanced cases. The policymakers and other stakeholders will have to tackle it in the upcoming years in order to develop a legal framework similar to those already governing other areas and aspects of the society [193].

V. Summary

This article has investigated the challenge of demographic bias in biometric systems. Following an overview of the topic and challenges associated therewith, a comprehensive survey of the literature on bias estimation and mitigation in biometric algorithms has been conducted. It has been found that demographic factors can have a large influence on various biometric algorithms and that current algorithms tend to exhibit some degree of bias w.r.t. certain demographic groups. Most effects are algorithm-dependent, but some consistent trends do also appear (as discussed in subsection III-C). Specifically, many studies point to a lower biometric performance for females and youngest subjects in biometric recognition systems, as well as lower classification accuracy for dark-skinned females in classification of demographic attributes from facial images. It should be noted that many of the studies conducted their experiments using relatively small datasets, which emphasises the need for large-scale studies. In general, a broad spectrum of open technical (and other) challenges exists in this field (see section IV).

Biased automated decision systems can be detrimental to their users, with issues ranging from simple inconveniences, through disadvantages, to lasting serious harms. This relevance notwithstanding, the topic of algorithmic fairness is still relatively new, with many unexplored areas and few legal and practical provisions in existence. Recently, a growing academic and media coverage has emerged, where the overwhelming consensus appears to be that such systems need to be properly assessed (e.g. through independent benchmarks), compelled to some degree of transparency, accountability, and explainability in addition to guaranteeing some fairness definitions. Furthermore, it appears that, in certain cases, legal provisions might need to be introduced to regulate these technologies.

Automatic decision systems (including biometrics) are experiencing a rapid technological progress, thus simultaneously holding a potential of beneficial and harmful applications, as well as unintentional discrimination. Zweig et al. [17] even argued that the issues (including, but not limited to bias and fairness) concerning algorithmic decision systems are directly related to the so-called “quality of democracy” measure of countries. As such, developing proper frameworks and rules for such technologies is a large challenge which
the policymakers and the society as a whole must face in the upcoming future [194], [195].

ACKNOWLEDGEMENTS

This research work has been funded by the German Federal Ministry of Education and Research and the Hessen State Ministry for Higher Education, Research and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE. A. Dantcheva was funded by the French Government (National Research Agency, ANR), under Grant ANR-17-CE39-0002.

REFERENCES

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TTS.2020.2992344, IEEE Transactions on Technology and Society


This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.


[121] P. Smith and K. Ricanek, “Mitigating algorithmic bias: Evolving an augmentation policy that is non-biasing,” in Winter Conference on