The Hitchhiker's Guide to Bias and Fairness in Facial Affective Signal Processing

Jiaee Cheong*‡ , Sinan Kalkan†, Hatice Gunes*

* University of Cambridge † Middle East Technical University ‡ The Alan Turing Institute

Abstract

Given the increasing prevalence of facial analysis technology, the problem of bias in these tools is now becoming an even greater source of concern. Several studies have highlighted the pervasiveness of such discrimination and many have sought to address the problem by proposing solutions to mitigate them. Despite these, to date, understanding, investigating and mitigating bias for facial affect analysis remain an understudied problem. In this work we aim to provide a guide by (i) providing an overview of the various definitions of bias and measures of fairness within the field of facial affective signal processing and (ii) categorising the algorithms and techniques that can be used to investigate and mitigate bias in facial affective signal processing. We present the opportunities and limitations within the current body of work, discuss the gathered findings and propose areas that call for further research.

Index Terms

Bias and Fairness, Facial Analysis, Facial Affect Analysis, Overview, Guide.

I. Introduction

Facial analysis, including identity recognition, facial attribute (e.g. age, gender, race) and affect prediction from facial images, has been widely studied in the literature, with increasing applications in various domains ranging from medicine, marketing, and surveillance [1]. With advances in machine learning, facial analysis is now dominantly performed using learning-based approaches. However, these data-driven, data-hungry, and black-box learning-based approaches make facial analysis biased and unfair towards certain demographic groups. Left unaddressed, bias can lead to concerning unfairness issues such

J. Cheong is supported by the Alan Turing Institute Doctoral Studentship.

S. Kalkan is supported by the BAGEP Award of the Science Academy, Turkey.

H. Gunes is funded by the EPSRC under grant ref. EP/R030782/1, and partially by the European Union's Horizon 2020 Research & Innovation programme project WorkingAge under grant agreement No.826232.

as misidentifying people of a certain race or gender at higher rates, misclassifying facial attributes much higher than any other demographic group [1], or predicting a higher probability of criminality for certain group of people [2].A biased depression analyser[1] or chronic pain detector[2] can have serious consequences if such technology is incorporated in the healthcare domain. Given the far-reaching real-life consequences that such threats pose, this issue has become a major concern. Within the public, this pressing issue has elicited widespread activism which eventually prompted governmental institutes such as the European Commission to set up a regulatory framework to address it [3].

Owing to its dire consequences, mitigating bias in learning-based approaches has attracted significant attention in the literature. Mitigation strategies attribute bias to the different stages of a learning-based approach, and therefore, they can be broadly analyzed in terms of the stage where they tackle bias: (i) Pre-processing (data-level) methods, (ii) In-processing (learning-model-level) methods or (iii) Post-processing (prediction-level) methods. Pre-processing methods, assuming that the data itself is biased, propose changing the amount of data instances (using over-sampling, under-sampling or data generation techniques) in favor of the demographic group affected by bias. In-processing methods, on the other hand, rely on modifying the algorithm to provide a fairer outcome. In contrast, post-processing methods propose adjusting the outputs of the learning-based model to achieve fairer predictions for all demographic groups.

Although several studies highlight bias and propose fairer solutions in face recognition, only a handful of studies have done the same for facial affect analysis [4], [5], [6], [7]. Facial affect analysis, despite having parallels to face recognition, bears inherent differences with the latter. Examples include it being more complex, subjective and ambiguous in terms of the nature of the data and its labels; incorporating significant discrepancy between lab-collected data and affect displayed in the real world; relying on data collected in context-free settings, which may be insufficient for recognition; lacking datasets for varied age and demographic groups; and requiring spatio-temporal data and annotations.

This dissimilarity calls for an informative investigation which is what we attempt in this paper. We present a hitchhiker's guide for understanding, detecting and mitigating "bias" for facial affect analysis by attempting to consolidate information across three distinctive fields: bias and fairness, facial affective signal processing, and machine learning. Our guide makes the following contributions. First, we provide an outline of the different concepts of bias and metrics of fairness relevant to facial affect signal processing. Second, we present an overview of the different types of bias mitigation strategies and discuss them from

[1] https://dl.acm.org/doi/abs/10.1145/3107990.3108004

[2] https://ieeexplore.ieee.org/abstract/document/7173007

an facial affect analysis perspective. Lastly, we discuss the challenges that remain to be addressed for fair facial affect analysis. Due to the phylogenetic nature of bias and fairness, we would like to emphasise that any attempt to shoe-horn a discourse such as this within a contrived space may risk rendering a multi-faceted topic into a simplistic caricature. Thus, we hope to highlight that this guide merely touches the tip of the iceberg. The included forms of bias and fairness within this guide are the ones which have been discussed or investigated within the existing facial affect literature. We will attempt to direct readers to other sources for more information on topics less extensively covered wherever possible. Bias and fairness can be analysed from several distinct angles and different interpretations will be useful to provide a more rounded analysis for the task at hand.

II. Bias and Fairness in Facial Affect Recognition

A. Facial Affective Signal Processing

There are different ways to distinguish one facial expression from another. The most common way is to describe an expression as discrete categories. Paul Ekman and his colleagues proposed that there are six basic emotion categories of facial expressions (i.e. happiness, surprise, fear, disgust, anger and sadness) each with particular configurations, recognised universally [8]. The basic emotion theory has been widely used for automatic analysis of facial expressions but has received criticism because basic emotions cannot explain the full range of facial expressions in naturalistic settings [9]. Therefore, a number of researchers advocated the use of dimensional description of human affect based on the hypothesis that each affective state represents a bipolar entity being part of the same continuum. The proposed polars are arousal (relaxed vs. aroused) and valence (pleasant vs. unpleasant) [9]. Another way to describe and analyse facial affect is by using the Facial Action Coding System (FACS), which is a taxonomy of human facial expressions in the form of Action Units (AUs) [8]. Action Units are the fundamental actions of individual muscles or groups of muscles. Since any facial expression results from the activation of a set of facial muscles, every possible facial expression can be comprehensively described as a combination of Action Units (AUs) – e.g., the facial expression of the happiness category can be described as a combination of pulling the lip corners up (AU 12) and raising the cheeks (AU 6).

Facial affective signal processing involves automatically analysing and predicting facial affect. As discussed above, depending on the method of distinguishing facial affect, this can be achieved by either training an algorithm to classify the facial expressions of emotion [5], predict the valence and the arousal value of the displayed facial expression or detect the activated facial action units [10], [11]. In this paper, we adopt a machine learning approach and assume that we are provided with a dataset D which is made up of pairs of $\{(\mathbf{x}_i, y_i)\}_i$ values where $\mathbf{x}_i \in X$ is a tensor representing information (e.g. facial image, health record, legal history) about an individual I and $y_i \in Y$ is an outcome (e.g. identity, age, emotion, facial action unit labels) that we wish to predict. In other words, we are interested in finding a predictor/mapping H with $H : X \to Y$. Each individual I is associated with a set of sensitive attributes $\{s_{j \in a}\}_a \subset S$ where a is e.g. *race* and $j \in \{$Caucasian, African-American, Asian$\}$. The sensitive attributes identified in the facial affect literature thus far are race, gender and age (readers can refer to the following papers for more information [5], [6], [7], [11]).The members of a minority group defined across a sensitive attribute (e.g., race = African-American) are usually protected by law and hence is often referred to as the protected group. Depending on the axis of sensitive attribute chosen, the remaining population are consequently referred to as the unprotected group within existing literature [12], [13]. Note that there are other attributes $\{z_{j \in a}\}_a \subset Z$ that are not sensitive.

B. Bias

An important step before addressing bias is identifying it and measuring whether our mitigation strategy provided fairer predictions. However, as research on bias in facial affect signal processing is still in its nascent stages, there is currently no consensus on how bias should be defined [7]. In fact, the term “bias” has been used to describe a wide range of concerns, even though each concern is unfair to the different groups in different ways and for different reasons. For instance, taking the example of race as a sensitive attribute, unequal representation in dataset, differences in performance accuracy, or higher prediction probability of prediction of an outcome are often all similarly described as “racial bias”. Given the large scope of discussion within this field, it would not be possible to thoroughly examine the nuanced formalisation of each definition here in this paper. Instead, we will highlight the common ways in which bias has been identified and analysed in the field of facial affective signal processing. In order to create a basis for the formulations that will follow, we use the following working definition for bias: Bias can be understood as ‘factors that lead to unintended consequences’ [3].

There are several commonly identified sources of bias within the field of machine learning and facial affective signal processing. One of the most pervasive types of bias is the *dataset bias*, i.e., the bias that is inherently present within the dataset, D [1], [2], [5], [11]. Dataset bias has been attributed to different factors affecting the dataset collection processes. Examples include sampling bias and data management bias. *Sampling bias* can typically be identified by comparing the statistical characteristics of the dataset

[3] https://arxiv.org/pdf/1901.10002.pdf

and the population it represents. It is manifested when the characteristics of the dataset population differs from that of the original target population. *Data Management bias* arises because incorrect data processing or handling results in a biased representation of the population. We would like to highlight that though the term “bias” has also been used to discuss other biases such as the dataset bias between lab-controlled datasets vs real-world data collections [14] and cross-dataset bias [15], we do not consider such examples as bias within this review. In the field of facial recognition in general, it is known that algorithms trained on lab-controlled datasets would perform much poorly for real-world data collections. Our focus in this paper instead is on the type of bias which will adversely impact subgroups of people.

One prevalent method of evaluating bias in facial affect signal processing is by measuring the tendency against a sensitive attribute $s_{i \in a}$ (e.g. Asian) that leads to worse H performance for the members of $s_{i \in a}$ compared to the members of $s_{j \in a}$ (e.g. Caucasian), $i \neq j$ [1], [5], [6], [11]. Another way of assessing bias is to leverage a counterfactual approach by examining how the affect recognition output varies in relation to a change in the sensitive attribute $s_{i \in a}$ [16]. Assuming all else is held constant, a facial affect recognition algorithm is considered biased if the output of the model $H(\mathbf{x}_i)$ is affected by a change in its sensitive attribute $s_{j \in a}$.

C. Fairness

TABLE I
AN OVERVIEW OF FAIRNESS DEFINITIONS.

Definitions of Fairness	**Description**
Group Fairness	Subjects in both the protected group p and unprotected group u have equal probability of being assigned to the positive predicted class. Several metrics of fairness are built on this definition of fairness.
Fairness Through Awareness	Any two individuals who are similar to a similarity metric defined for a particular task should receive a similar outcome.
Fairness Through Unawareness	An algorithm is fair as long as any sensitive attributes S are not explicitly used in the decision-making process.
Counterfactual Fairness	An outcome is fair towards an individual if it remains the same in a counterfactual world where the individual belonged to a different demographic group.
Fairness in Relational Domains	This notion of fairness captures the relational structure in a domain by taking into account the social, organizational and other connections between individuals in addition to the attribute of an individual.

Similar to bias, to date, there is no universal definition of fairness. Each different machine learning task (e.g., facial affect recognition, credit worthiness prediction) will result in very different datasets, problem formulations and end-goals. It is difficult to establish an all-inclusive valid definition of fairness due to the sheer insurmountability of accounting for every different factors, scenarios and tasks. Although it may seem intuitive to associate a decrease in bias with an increase in fairness, this may not be always true. The link between bias and fairness is largely dependent on the definition of bias and the criteria for fairness. Some of the representative types of fairness within the affect recognition literature and their definitions are provided in Table I. Due to the limited scope of this paper, we are only able to focus on a facile aggregation of the types of fairness relevant to affective signal processing. Further examples and a more comprehensive discussion of the other forms of fairness can be found in [12] and [13]. In addition, although the metrics and definitions provided largely pertain to binary classifications, these can be easily extended to multi-class classification problems [5].

In facial affective signal processing, researchers generally associate a decrease in bias as an increase in fairness [5], [6], [11]. In order to achieve greater fairness, we can either ensure a fairer dataset or a fairer affect analysis output. To address the former, researchers typically evaluate the proportion of each subgroup $s_{i \in a}$ in the dataset D and consider the dataset D less biased or fairer if this proportion is the same for all members (e.g. Caucasian, African-American, Asian) of attribute a (e.g. race). To address the later, researchers generally measure the fairness of a facial affect analysis algorithm by comparing the performance of the algorithm among the different sensitive groups. We can say that the outcomes are fair if they are similar in terms of some performance metric(s), e.g. True Positive rate, False Positive rate, for different groups. More formally, denoting such a performance metric m on demographic group $s_{i \in a}$ by $d_m(s_{i \in a})$, fairness can be evaluated comparing $d_m(s_{i \in a})$ and $d_m(s_{j \in a})$. Different choices for $d_m(\cdot)$ lead to different fairness metrics and characteristics [5], [11], [12] – see Table II for some examples. For instance, for a = gender, and m = accuracy, we would compare $d_m(s_{male})$ and $d_m(s_{female})$ to evaluate whether the method $H : X \to Y$ satisfies the equal accuracy metric of fairness across gender. We can investigate whether the accuracy of the affect recognition method $H : X \to Y$ is consistent across the different subgroups $s_{i \in a}$ for all $i \in a$. A greater consistency in accuracy across s_{male} and s_{female} is associated with a greater fairness [5], [6], [11].

In addition, a hitchhiker should also be aware of the fairness-accuracy trade-off. In general, if the model generalises well and is fairer across different subgroups, it is likely to be less accurate than a model which is only tuned to achieve a high performance [11]. For every task or application, its resulting adverse impact is different. The cost of misclassification (especially those from the minority group) needs to be carefully considered and taken into account. Hence, it is necessary for each practitioner to be aware of this trade-off and find the right fairness-accuracy balancing point for the task at hand.

There are several tools which the practitioner can make use of. Open-source libraries such as *Aequitas* [17] offer an auditing toolkit which calculates a list of performance-based bias scores according to different bias metrics. In addition, Aequitas also provides a decision tree to help users identify the most appropriate metric to use depending on the task at hand. *Fairness Measures* library [18] offers a similar functionality and also includes additional metrics and some datasets to experiment with. *FairTest* [19], on the other hand, approaches the same task by quantifying the associations between predicted labels and protected attributes. In addition, it also identifies regions of the input space where the classifier might incur unusually high error rates and includes a collection of datasets to allow for testing.

TABLE II
OVERVIEW OF METRICS OF FAIRNESS.

Metrics of Fairness	**Description**	**Formulation**
Equalised Odds	Protected group p and unprotected group u should have equal rates for true positives and false positives.	$\frac{TP_p}{FP_p} = \frac{TP_u}{FP_u}$
Equal Accuracy	Protected group p and unprotected group u should have equal rates of accuracy.	$\frac{(TP_p+TN_p)}{(TP_p+TN_p+FP_p+FN_p)} = \frac{(TP_u+TN_u)}{TP_u+TN_u+FP_u+FN_u}$
Equal Opportunity	Protected group p and unprotected group u should have equal true positive rates	$\frac{TP_p}{(TP_p+TN_p+FP_p+FN_p)} = \frac{TP_u}{TP_u+TN_u+FP_u+FN_u}$
Demographic Parity (aka Statistical Parity)	Likelihood of a positive outcome should be the same regardless of whether the person is in the protected group.	$P(y\|s_i^a = p) = P(y\|s_j^a = u)$
Treatment Equality	Treatment equality is achieved when the ratio of false negatives and false positives is the same for both p and u.	$\frac{FN_p}{FP_p} = \frac{FN_u}{FP_u}$
Test Fairness	For any predicted probability score $\hat{y}$, people in both p and u must have equal probability of correctly belonging to the positive class.	$P(y = 1\|\hat{y}, s_i^a = p) = P(y = 1\|\hat{y}, s_j^a = u)$
Conditional Statistical Parity	People in p and u should have equal probability of being assigned to a positive outcome given a set of legitimate factors L.	$P(y = 1\|L, s_i^a = p) = P(y = 1\|L, s_j^a = u)$

Potential legitimate factors would be how quickly lip corners are pulled up, which specific facial action units are activated and what their intensities are. True positive (TP): predicted and actual outcomes both positive. False positive (FP): predicted positive but actual negative. False negative (FN): predicted negative class but actual positive. True negative (TN): predicted and actual both negative. Although the metrics and definitions provided largely pertain to binary classifications, these can be easily extended to multi-class classification problems. Readers can refer to [13] for a more detailed review.

III. Mitigating Bias

There exist a variety of methods to mitigate the effect of bias and increase fairness. These methods attribute and address bias at different stages of the processing pipeline, as illustrated in Figure 1. Based on the stage where they address bias, the mitigation methods can be broken down into (i) pre-processing methods which tackle the bias problem at the data level, (ii) in-processing methods which mitigate the bias problem at the model training level and (iii) post-processing methods which address the problem at the output level. Figure 2 provides an overview and a taxonomy of the tools available to a hitchhiker. Readers can refer to [5], [6], [11] for further information.

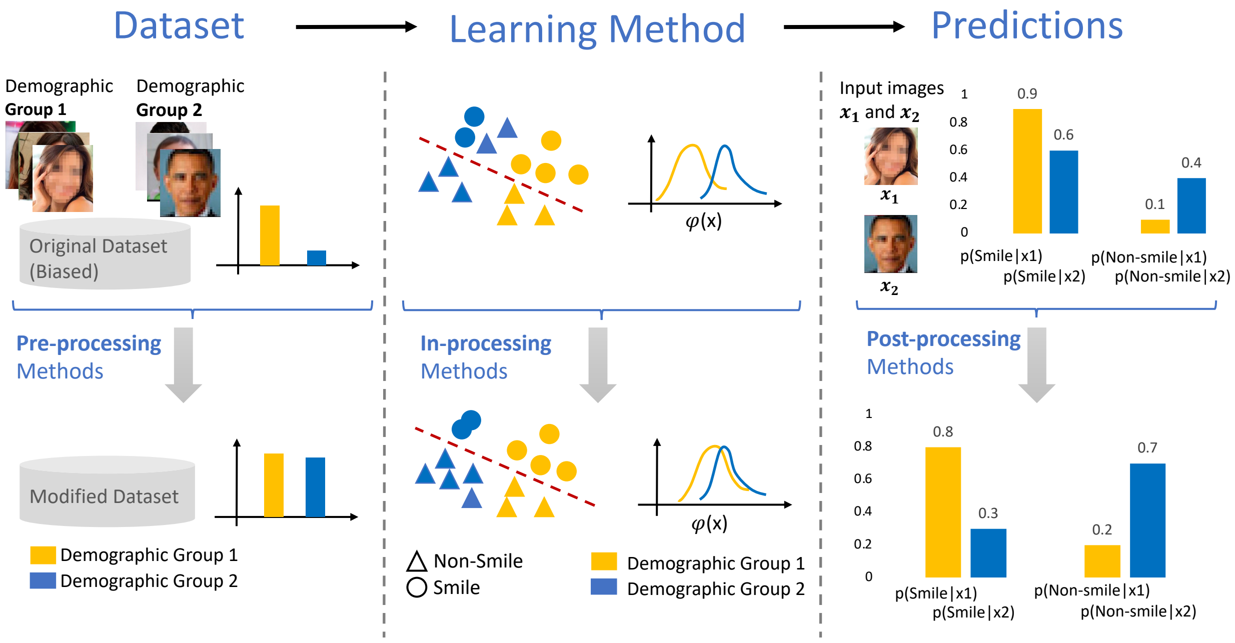

Fig. 1. The three different strategies tackle bias at the different stages of a facial analysis system. The pre-processing methods address bias by changing the amount of data and aim to solve the issue irrespective of the learning method. The in-processing methods, aim to achieve fairness by providing the same feature representations or objective scores for the different demographic groups. In contrast, the post-processing methods adjust the outputs of the learning model to provide fairer predictions.

A. *Pre-processing (Data-level) Methods*

Data imbalance is a persistent problem in the field of machine learning – a dataset is considered balanced if the proportion of individuals across the different semantic categories (e.g., gender, age group and skin colour) are roughly equal. Studies have highlighted how data imbalance results in biased predictions and this is true for both face recognition and facial affect analysis algorithms [20]. Data-level approaches typically attempt to mitigate bias by (i) creating a new dataset, (ii) leveraging on resampling methods, (iii) data augmentation or (iv) synthetic data generation and (v) fairness through unawareness. Note that, at times, instead of using the terminology "fairness through unawareness", certain papers have opted to refer to it as "feature transformation" instead.

1) Fairness via Creating Balanced Dataset: In this method, bias is mitigated by creating a new dataset D_n ensuring an equal proportion for each $s_{i \in a}$. However, creating a new balanced dataset with an equal proportion of individuals across all categories of sensitive attributes is a challenging task. In addition, the collected dataset has to be sufficiently large as face analysis and recognition algorithms typically do not learn nor perform well when the sample-sizes are too small. This is labour intensive as the datasets are typically manually collected and attributes such as age and gender has to be labelled by hand [20]. Nonetheless, creating a dataset from scratch will allow researchers to have greater control over the sample population statistics. The problem of dataset bias can, hence, be significantly reduced by ensuring that the collected dataset has an equal proportion of each population subgroup. The example given here describes an approach which advocates for having equal subgroup representations in order to ensure that the model learns equally well from all subgroups. Another definition of curating a balanced dataset is to have a dataset which is a reflection of the population statistic. Both approaches will lead to very different outcomes and referring to both as dataset bias might be construed as a misnomer.

2) Fairness via Re-sampling: There are two main resampling methods to consider: Down-sampling and over-sampling. Down-sampling / dropping is simply the process of sampling a smaller, balanced subset of the original data D_o as the training data. This can be achieved by discarding some $\mathbf{x}$'s with the over-represented sensitive attribute $s_{i \in a}$ from D_o such that the eventual training set D_{tr} has an equal proportion for each $s_{j \in a}$. This solution is able to achieve a fairer dataset despite the information loss it introduces. In fact, in addition to contributing to a higher classification accuracy, Xu and Jin [21] have shown that a down-sampling method can also reduce the variance in the different poses and expressions of the same face. In contrast, *over-sampling* increases the training set D_{tr} by duplicating instances of the minority group $s_{i \in a}$ to achieve a balanced dataset. The selection of the instances to be duplicated is typically done at random.

3) Fairness via Data Augmentation: Data augmentation is a synthetic data-generation technique to synthesize more training data. Similar to the re-sampling methods discussed previously, data augmentation combats the problem of dataset bias by training the algorithm on a training set that is less biased. In contrast to oversampling, this approach does not duplicate instances. Instead, the additional samples generated using data augmentation typically involve a certain degree of input transformation. In other words, re-sampling methods end up with a new D_{tr} with X duplicated or subsampled. However, with data augmentation, D_{tr} includes samples that are not in D_o.

Data augmentation can be done in several different ways. Simpler data augmentation methods include image flip, rotation, colour augmentation and random crops [5]. Such methods introduce more training data which, therefore, increases the robustness of a model and prevent the model from over-fitting to the over-represented demographic groups [5]. However, such simple augmentations may not provide the needed intra-class variation to produce distinct samples to re-balance the dataset. More complex methods include the well-known Synthetic Minority Over-sampling Technique (SMOTE) [22] which can generate synthetic instances in the vicinity of the minority group via the k-nearest-neighbour method. Overall, classifiers trained on augmented data have shown to exhibit positive improvement in terms of fairer classification performance [5].

4) Fairness via Synthetic Data Generation: The state-of-the-art in synthetic data generation has progressed significantly in recent years and we can now generate sophisticated, high-quality, synthetic images. This is achieved using generative models such as variational autoencoders (VAEs) or generative adversarial networks (GANs) where a generator (G) is used to map a latent vector $\mathbf{z}$ to an image $\hat{\mathbf{x}} = \mathrm{G}(\mathbf{z})$. VAEs and GANs are trained such that a generated sample $\hat{\mathbf{x}}$ is not distinguishable from $\mathbf{x} \in D_o$ with respect to some pre-defined criteria. For mitigating bias, the generator (G) is used to produce a new synthetic dataset D_{syn} to augment the training set and compensate for the underrepresented subgroups. Note that the synthetic dataset D_{syn} contains samples that are different from the original dataset D_o. The affect analysis algorithm can then be trained on the augmented balanced data to produce less biased predictions. Ngxande et al. [4] used synthetic data generation to reduce bias in driver drowsiness detection by defining a mini-max objective function which accounts for the expected log-likelihood that the sample $\mathbf{x}$ is from the data and the expected log-likelihood that the sample $\mathbf{x}$ is not from the generator (G).

5) Fairness through Unawareness: Fairness through unawareness, also known as feature transformation in the literature, is another pre-processing approach that can be used to mitigate dataset bias. It simply excludes the features related to the sensitive attributes S within the training dataset D_{tr}. The algorithm is trained to map $H : X \rightarrow Y$ with only the remaining features. This method assumes that the remaining features in the input do not carry any residuals of information about the sensitive attributes which may be impractical in several real-life tasks. If $\mathbf{x}$ is a facial image, fairness through unawareness might not be applicable as a pre-processing technique since we may not be able to ascribe sensitive attributes to pixels. However, if $\mathbf{x}$ includes sensitive features, such as gender and race, in addition to a facial image, fairness through unawareness can be employed. In other words, if there are additional attributes that may include sensitive information provided as variables during the training stage, fairness through unawareness might be a suitable method. To use the example of race, this simply means removing the variable race from the training dataset when training the algorithm. However, there might still be other features correlated with race which may undermine the effectiveness of this method.

Discussion

Though relatively successful, the pre-processing techniques have certain drawbacks. For example, down-sampling might lose critical samples and over-sampling might cause over-fitting since exact copies of the minority class are replicated randomly. It is possible to combine the down- and over-sampling procedures. Though such methods have performed well, the increase in procedural steps also inadvertently increases the computational cost necessary for data pre-processing and model training. In addition, popular methods such as SMOTE are shown to lead to over-generalization and are largely ineffective in dealing with highly imbalanced datasets. The re-sampled data-points may even amplify the bias in the data if the original dataset contains too few instances of the minority group for it to be sufficiently representative.

Although generative models are successful at generating realistic synthetic samples for targeted data augmentation for a facial affect analysis task [4], we still have to recognise that generative models are not without its drawbacks. These methods typically generate synthetic data that matches the distribution of the original training set. The synthetic data samples generated by an algorithm trained on biased data will most likely inherit the bias from the original dataset. Unless special care is taken in controlling the attributes of the generated samples, they might just propagate the existing biases of the original data the generative model was trained on. Moreover, such methods can be difficult to train to obtain realistic and diverse synthetic samples, and they are computationally expensive. In addition, besides the fairness via unawareness method, the other methods in the pre-processing stage largely ensure that the dataset is balanced but might not necessarily satisfy any of the fairness definitions highlighted in Table I.

B. In-Processing Methods

In contrast to the data-level approaches, in-processing methods are mainly algorithm- or model-level approaches. Such approaches typically attempt to mitigate bias by augmenting the learning procedure to improve the sensitivity of the model towards under-represented groups. The objective of doing so is to obtain better prediction results for the minority group without affecting the performance on the majority class. In-processing methods can be divided into three main groups: (i) cost-sensitive learning techniques, (ii) domain adaptation methods and (iii) disentanglement approaches. Note that, at times, the naming convention varies between practitioners. For instance, several papers do not use the term disentanglement and merely refer to the methods as “adversarial”.

1) Fairness via Cost Sensitive Learning: Cost sensitive learning is a method that relies on penalizing prediction errors $\ell(\mathbf{x})$ for different demographic groups by weighting the prediction errors, i.e. $w_{s_{i\in a}} \cdot \ell(\mathbf{x})$ where $\mathbf{x} \in s_{i\in a}$. The weight $w_{s_{i\in a}}$ can be determined based on the overall performance of the model for the demographic group $s_{i\in a}$ or the number of instances in each demographic group. For example, if the ratio of images belonging to female and male groups is $1 : 3$, w_{female} can selected such that $w_{female} = 3 \cdot w_{male}$.

The aim in cost sensitive learning is to guide the classifier to place more emphasis on minority samples in order to combat the biases within the training dataset. For instance, it is viable to give greater importance to the minority class samples via the setting of weights. By setting a higher penalty on misclassifying a minority class, the classifying algorithm can be trained to be more attuned towards the minority classes than they would otherwise be. The classifying algorithm is therefore better able to deal with the initial dataset bias and produce a fairer outcome. There is likely to be a fairness-accuracy trade-off involved as improving the accuracy of the minority group might reduce the accuracy for the majority group. The weights should therefore be decided with this trade-off in mind in order to find a good balance between fairness and accuracy for the task at hand.

2) Fairness via Domain Adaptation: Domain adaptation is another promising method to address bias. Such solutions are especially relevant in practical scenarios where the target face database often has a different distribution from that of the source training dataset. For instance, several works have shown that models trained on dataset of another age group do not generalize well on children [6] and elders [23] for facial affect recognition tasks. This is because the majority of the facial affect datasets used for algorithm training contain facial affect displays of young and middle-aged adults. Facial affect data from older adults, adolescents and children is relatively less accessible and available [23]. This in turn results in algorithms which are relatively less sensitive to the variations due to age.

In order to mitigate the bias arising from such scenarios, domain adaptation addresses the discrepancy between probability distributions $P_{\mathbf{x}\in s_i}(\mathbf{x})$ and $P_{\mathbf{x}\in s_j}(\mathbf{x})$ where s_i and s_j are respectively under-represented and over-represented groups. Domain adaptation methods generally try to reduce the difference between the features in a deep network for under-represented or over-represented groups, namely, $\phi_{\mathbf{x}\in s_i}(\mathbf{x})$ and $\phi_{\mathbf{x}\in s_j}(\mathbf{x})$. This can be achieved either directly, by minimizing a cost term based on the difference between $\phi_{\mathbf{x}\in s_i}(\mathbf{x})$ and $\phi_{\mathbf{x}\in s_j}(\mathbf{x})$, or by enforcing a prior on these features, e.g. a Gaussian.

Many adaptations of such strategies have been employed for bias mitigation in the literature. For example, mutual information between the features of Caucasian and non-Caucasian samples can be maximized to obtain the same feature distributions for all demographic groups [24]. In fact, a recent work extended the domain adaptation approach using a continual-learning mechanism and evaluated it on two different facial affect recognition tasks: facial expression classification and action unit detection [11]. Experimental results indicated that this method not only achieved a high performance accuracy score, but also managed to maintain fairness across the sensitive attribute-based group splits.

3) Fairness via Disentanglement: Disentanglement methods attempt to *separate* representations belonging to different attributes. In the case of bias mitigation, disentanglement is used to remove the sensitive demographic features from $\phi_{\mathbf{x}\in s_i}(\mathbf{x})$ and perform classification without these sensitive features. In a sense, this is a form of *fairness through unawareness* but at the model-level. This method has proven to be quite effective in mitigating bias in facial affect recognition. Xu et al. [5] compared three different bias mitigation approaches, namely, a baseline, an attribute-aware and a disentangled approach on two well-known datasets, RAF-DB and CelebA. Their results indicated that the disentangled approach is the best for mitigating demographic bias in facial affect recognition.

A popular disentanglement strategy is *adversarial learning*. A typical application of this incorporates an adversarial, sensitive-attribute predictor, in addition to the original prediction task, y. The goal of the learner is to obtain a representation $\phi(\mathbf{x})$ from which the adversarial predictor cannot successfully predict the attribute s_i^a but the original prediction, y, can be performed.

Discussion

In contrast with pre- and post-processing methods, in-processing methods can be significantly more effective; however, they are a lot more varied, diverse and harder to implement without prior knowledge of the algorithms' assumptions. Among the in-processing methods, cost-sensitive learning is the easiest to employ and commonly used for quick mitigation results. However, a strategy that solely relies on the number of samples or their prediction errors can be insufficient for removing bias entirely.Moreover, as we are making direct changes to the actual prediction algorithm, this will innately result in fundamental changes and may even introduce unwanted consequences. For instance, since men and women may express their emotions differently, reducing the feature difference between both genders may result in the emotion-discriminative information being reduced. As a result, this may give rise to an increase in fairness at the expense of a decrease in classification accuracy.This is one manifestation of the fairness-accuracy trade-off. In addition, many of the algorithms do not fall neatly into one category and often overlap or are somewhat related to one another. In fact, many of the methods are *hybrid models*. These models consist of multiple techniques which can be combined in a sequential, parallel or even multi-party system. For instance, Li and Deng [14] proposed a framework which combines cost sensitive learning and domain adaptation for fairer facial expression recognition.

C. *Post-Processing Methods*

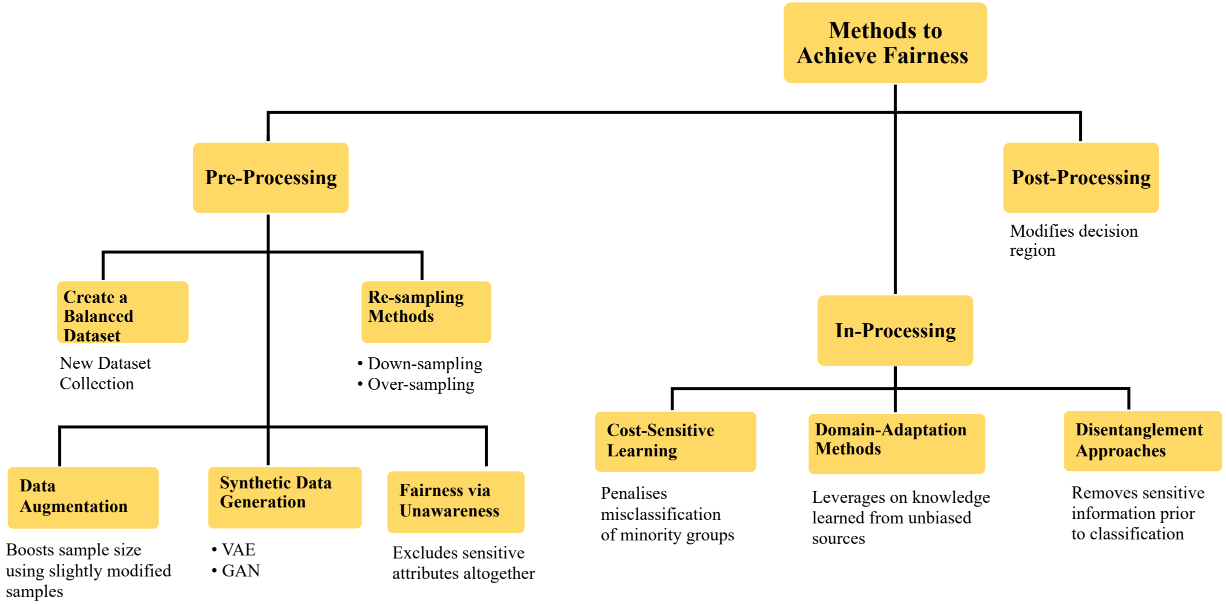
Methods to Achieve Fairness
Pre-Processing
Post-Processing
Modifies decision region
Create a Balanced Dataset
New Dataset Collection
Re-sampling Methods
• Down-sampling
• Over-sampling
In-Processing
Data Augmentation
Boosts sample size using slightly modified samples
Synthetic Data Generation
• VAE
• GAN
Fairness via Unawareness
Excludes sensitive attributes altogether
Cost-Sensitive Learning
Penalises misclassification of minority groups
Domain-Adaptation Methods
Leverages on knowledge learned from unbiased sources
Disentanglement Approaches
Removes sensitive information prior to classification

Fig. 2. Taxonomy of Bias Mitigation Methods.

Post-processing methods attempt to mitigate bias after the algorithm training and classification process is complete. In general, these approaches correct the bias in the algorithm by modifying the outputs of the algorithm directly to achieve a fairer output score. An example of this is the Multiaccuracy-Boost algorithm where the algorithm serves as an “auditer”. The auditor attempts to identify sub-populations where the original classifier systematically erred and iteratively post-processes the classifier until it achieves an unbiased accuracy on each subgroup [25]. Another example is the pioneering study conducted by Howard et al. [6] which investigated how using a cloud-based emotion recognition algorithm applied to children’s facial expression images can be skewed when performing recognition on the data of that minority class. To remedy this, they proposed a post-hoc hierarchical approach combining outputs from the cloud-based emotion recognition algorithm with a specialized learner.

The more integrated efforts attempt to bridge bias detection and mitigation in order to provide a more comprehensive solution. Themis-ML is such an example [26]. This repository provides a suite of fairness metrics similar to those discussed above and bias mitigation algorithms such as the additive counter-factually fair estimator and re-labelling technique to mitigate the bias detected. Fairness Comparison is another example [27]. In addition to offering a different set of bias mitigation methods, it primarily differs from Themis-ML by allowing the different bias metrics and algorithms to be methodologically compared and the addition of supplementary algorithms and datasets to its platform. Besides the efforts from academia, commercial organisations such as IBM also provided a comprehensive set of fairness metrics, explanations, and algorithms to mitigate bias on both the dataset and algorithmic levels [28]. The significant contribution of such commercial efforts chiefly lies in their architectural design which makes it more user-friendly and relevant towards non-academic practitioners within the industry.

Discussion

There are comparatively fewer literature on post-processing methods in comparison with pre-processing and in-processing methods. In general, post-processing methods are non-restrictive and can be used on almost any family of algorithms. This is especially desirable when we do not have access to the inner workings of the algorithm or we are unable to modify or re-train it to make it more fair. The additional key benefit is that there are more tools and packages available for post-processing methods [17], [26], [27]. Such tools could typically calculate several different metrics of fairness from Table II. This makes it easier for non-Machine Learning experts and general practitioners to understand, address and mitigate bias by customising their methods to achieve the most appropriate definition of fairness illustrated in Table I. Such tools thus make it easier for the end-user to provide a more tailored solution to the problem at hand and facilitate the process of ensuring any research developed to be put into practice. This is highly important as ultimately, this is the eventual goal of the research community: To ensure greater fairness within the algorithms deployed. Post-processing methods, therefore, provide the platform to facilitate the transition of fairness research in academia to real-world algorithm deployment.

IV. Discussion

In the previous sections, we have attempted to provide a systematic overview of bias mitigation methods available to practitioners within this field. In the field of facial affect recognition, although a systematic analysis of bias and the investigation of mitigation strategies are still in their infancy [5], [7], we have highlighted specific implementations of such methods on facial affect recognition tasks wherever possible. If not, we attempted to highlight implementations within the field of facial recognition which could be incorporated into facial affect recognition. We now proceed to discuss the challenges identified and provide insights for future work for bias mitigation in facial affect signal processing.

A key challenge in affect recognition is the difficulty to represent and label affect objectively. As discussed, bias and fairness are difficult concepts to formalise quantitatively and there is a lack of consensus on how it should be defined in the current literature. The same can be said of how emotions should be represented. Affect recognition deals with complex and at times subjective concepts - i.e., the

TABLE III
OVERALL COMPARISON OF THE THREE CATEGORIES OF METHODS.

Method	**Pros**	**Cons**
Pre-processing	✓ Can be applied before most algorithms. ✓ Straightforward implementation.	⚠ Limited efficacy in terms of bias mitigation.
In-Processing	✓ Streamlines the process as it does not require an additional pre-processing step. ✓ Performs best on accuracy and fairness measures. ✓ Provides the most flexibility for determining the accuracy-fairness trade-off.	⚠ Algorithm trained is task-specific and cannot be deployed on another task. ⚠ Need to re-train the algorithm, which may not be possible in many scenarios. ⚠ Likely to incur higher time and computational cost. ⚠ Harder to implement without sufficient knowledge.
Post-processing	✓ No need to re-train classifying algorithm. ✓ Can be applied after most algorithms. ✓ Does not require access to the raw data.	⚠ Black-box tools might not be able to provide good interpretability and explainability. ⚠ Lack the flexibility of determining the accuracy-fairness trade-off.

Practitioners may need to explore multiple methods at all three levels to achieve fairer decision-making especially for high-stakes situations.

labels used for training and testing are somewhat subjective and at times ambiguous as assessing both one's own and another person's affect are challenging tasks, and mapping these unambiguously to a single category or point in space is not straightforward [29]. As a result, there is currently a bias towards the "labelled" emotional states such as anger and happiness as compared to other forms of affect such as "annoyance". Algorithms trained on existing datasets will therefore be more prone towards categorising affect into the recognised category even though the predicted emotion category might not be the best fit. There is debate within the community on whether this is the best way forward or whether other metrics such as representing emotions on a continuous scale would be more appropriate or fairer approach instead.

Another challenge is that the lab versus real-world discrepancy becomes even more pronounced for affective computing as facial affect displayed within the lab might be less natural or more exaggerated as compared to real-world situations. For face recognition, lab-trained algorithms might struggle with real-world data due to differences in pose and illumination. These challenges also exist for facial affect recognition, in addition to other challenges such as context that involves a target user population and a target application domain which cannot be easily replicated within the lab. As a result, algorithms trained on expressions collected in a context-free environment are unlikely to generalise well and may be biased towards affect that are collected in context-free settings. Moreover, people are more likely to regulate their display of affect in lab settings. How these aspects would impact bias and fairness related issues in facial affect recognition needs further investigation.

Besides, the majority of the facial affect datasets used for training automatic affect recognisers contain static facial images. In line with this, current research on bias and fairness in facial affect recognition has started with investigations on static image datasets. However, understanding and correctly interpreting user affect requires data, labels and methodologies that consider the (spatio-)temporal aspects [30] which may also show differences in lab vs. real-world settings in terms of timing of the nonverbal behaviour (e.g., the temporal evolution of a fake smile is different from that of a genuine smile). In order to see the impact of such factors on the fairness for facial affect recognition, they need to be investigated with relevant experiments and compared to using only static image datasets.

In addition, the majority of the existing facial affect datasets that are made publicly available for research purposes do not contain information regarding the sensitive attributes, making it difficult to assess bias, let alone mitigate it. Further research should be undertaken to understand which mitigation strategies are needed when considering the (spatio-)temporal aspects of the affective displays as well as context-free vs. context-dependent settings.

V. Recommendations and Concluding Remarks

In this review, we investigated the topic of bias and fairness in facial affect recognition systems. We did this by formalising the fundamental components of bias and fairness relevant to facial affect signal processing, investigating the progress made and highlighting the challenges present within this area of research. In Table III, we provide an overview of pros and cons for pre-processing, in-processing and post-processing bias mitigation methods.

Additionally, in Figure 3, we provide a modified version of the flowchart found in [28] to illustrate the mechanism, the metrics and the pipeline that can be adopted for investigating bias and obtaining fairer facial affect recognisers. The paths are options that can be taken and do not constitute a formal recommendation. The guide adopts a holistic approach along several dimensions - i.e., data is loaded, and then transformed into a (fairer) dataset using a fair pre-processing algorithm, and a classifier is learned from this transformed dataset. Predictions are then obtained and metrics are calculated on both the original and the transformed dataset as well as between the transformed dataset and the eventual dataset with the predicted outcome.

Furthermore, we would like to emphasise that the guide that we have provided lacks efficacy if it is not considered in a holistic manner alongside the substantial discourse on the ethical implications of such innovation. As it stands, it is widely perceived that efforts from academia and big tech companies,

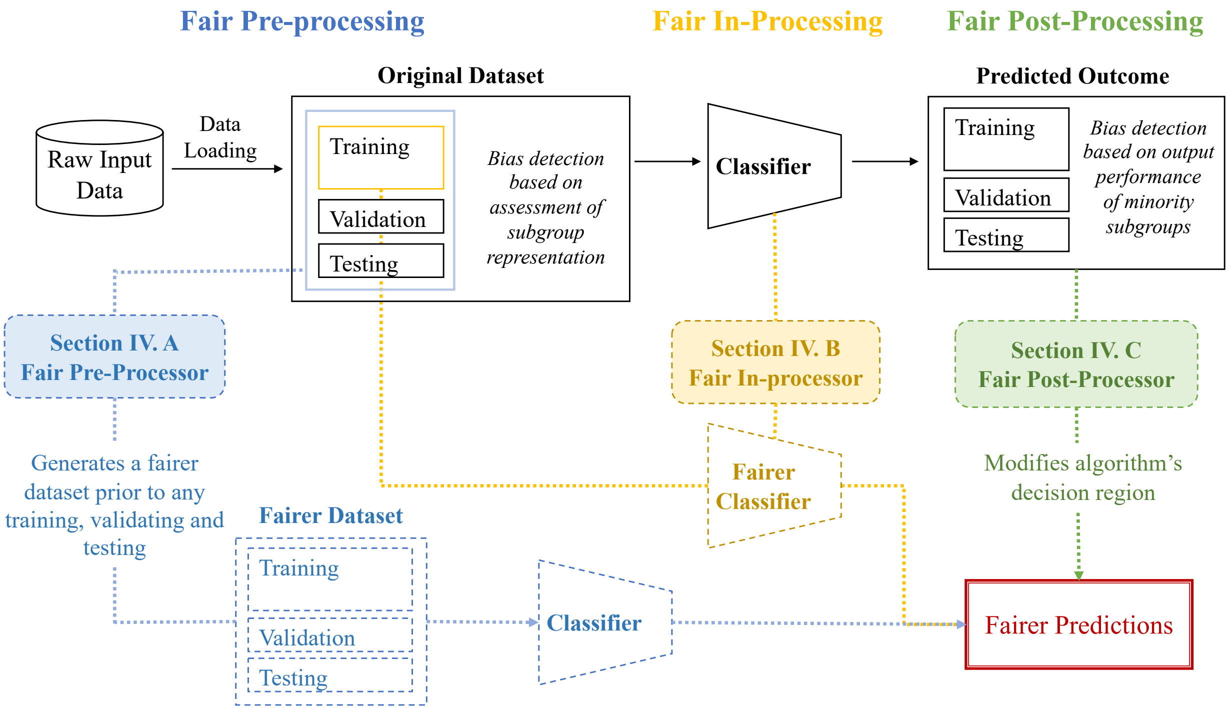
Fair Pre-processing
Fair In-Processing
Fair Post-Processing
Original Dataset
Predicted Outcome
Raw Input Data
Data Loading
Training
Validation
Testing
Bias detection based on assessment of subgroup representation
Classifier
Training
Validation
Testing
Bias detection based on output performance of minority subgroups
Section IV. A Fair Pre-Processor
Section IV. B Fair In-processor
Section IV. C Fair Post-Processor
Generates a fairer dataset prior to any training, validating and testing
Fairer Classifier
Modifies algorithm's decision region
Fairer Dataset
Training
Validation
Testing
Classifier
Fairer Predictions

Fig. 3. The Hitchhiker's Guide to the Overall Process of Bias and Fairness in Facial Affective Signal Processing: Mechanism, Metrics and Pipeline. This is a modified version of the flowchart found in [28] to provide an overview of the options available to a hitchhiker. Note that the paths are options that can be taken and do not constitute a formal recommendation. The hitchhiker is encouraged to experiment with a myriad of approaches and leverage any permutation of strategies in order to maximise bias reduction without overly sacrificing accuracy for the task at hand.

though gaining traction, are still relatively inadequate [2] and that ethical reforms are largely reliant on regulation, governmental guidelines and legislation to enforce responsible innovation [3]. Given the repercussions discussed previously, as a field, we need to start contending with how our work might cause or be implicated in harm and consider to what extent is the research community responsible for making sure that the technology we are developing is ethical. The panel discussion on the *Potential Misuse of the Affective Computing Science and Technology* at ACII 2019, and ACII 2021's theme of *Ethical Affective Computing* are good starting points for the community for considering ethical uses, fairness and bias in emotional AI. Another highly important and difficult question that plagues the community is how fair do our algorithms need to be before it is deemed acceptable? Though the governmental and legal guidelines in place might not be the best arbitrator for such discourse, they represent a step forward towards ensuring that the harm arising from bias in high-stakes situations such as "automated video interviews" are somewhat curtailed.

Moreover, the ethics of these problems are not solely restricted to bias and fairness. There is also a trade-off between privacy and efficacy. Having access to more data would typically result in a more efficacious (but not necessarily fairer) solution. Coming back to the example of the driver drowsiness detection [4], would it be worth collecting more data to improve the fairness of the algorithm at the cost of individual privacy? This is not an easy question to answer. Given the complexity surrounding these issues, to foist a conclusive form of how fairness should be achieved may risk incurring the epistemic fallacy that the epistemological priority of definitional knowledge would be a panacea for the problem of bias. It also risks masquerading as a red herring to the more fundamental forces at play: that prescribing legalistic injunctions about what constitutes fairness may very well exacerbate the bias that we wish to avoid. As such, we we would like to emphasise that the goal of ethical, responsible and fair innovation cannot be achieved through exogenous guidelines (legal or otherwise) alone. This guide is not intended to be canonical nor axiomatic. Our hope is that it will serve as a starter's guide for a hitchhiker to navigate through the kaleidoscope of definitions, methodologies, tools and concepts as they journey towards developing fairer solutions for all.

REFERENCES

[1] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proc. ACM Conf. Fairness Accountability Transparency*, Jan. 2018, pp. 77–91.

[2] M. Garcia, "Racist In The Machine: The Disturbing Implications Of Algorithmic Bias," *World Policy Journal*, vol. 33, no. 4, pp. 111–117, 2017.

[3] E. Commission, "White paper on artificial intelligence – a european approach to excellence and trust," *European Commission*, 2020.

[4] M. Ngxande, J. Tapamo, and M. Burke, "Bias remediation in driver drowsiness detection systems using generative adversarial networks," *IEEE Access*, vol. 8, pp. 55 592–55 601, 2020.

[5] T. Xu, J. White, S. Kalkan, and H. Gunes, "Investigating bias and fairness in facial expression recognition," in *ECCV2020 Workshop: ChaLearn Looking at People workshop ECCV: Fair Face Recognition and Analysis*, 2020.

[6] A. Howard, C. Zhang, and E. Horvitz, "Addressing bias in machine learning algorithms: A pilot study on emotion recognition for intelligent systems," in *2017 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO)*, 2017.

[7] A. Domnich and G. Anbarjafari, "Responsible ai: Gender bias assessment in emotion recognition," *arXiv preprint arXiv:2103.11436*, 2021.

[8] R. Ekman, *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.

[9] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, pp. 1161–1178, 1980.

[10] J. Deuschel, B. Finzel, and I. Rieger, "Uncovering the bias in facial expressions," *arXiv preprint arXiv:2011.11311*, 2020.

[11] N. Churamani, O. Kara, and H. Gunes, "Domain-incremental continual learning for mitigating bias in facial expression and action unit recognition," *arXiv preprint arXiv:2103.08637*, 2021.

[12] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *arXiv preprint arXiv:1908.09635*, 2019.

[13] S. Verma and J. Rubin, "Fairness definitions explained," in *Proc. ACM Int. Workshop Softw. Fairness (FairWare)*, 2018.

[14] S. Li and W. Deng, "A Deeper Look at Facial Expression Dataset Bias," in *IEEE Tran. on Affective Computing*, 2020.

[15] B. Han, W.-H. Yun, J.-H. Yoo, and W. H. Kim, "Toward unbiased facial expression recognition in the wild via cross-dataset adaptation," *IEEE Access*, vol. 8, pp. 159 172–159 181, 2020.

[16] E. Denton, B. Hutchinson, M. Mitchell, and T. Gebru, "Detecting bias with generative counterfactual face attribute augmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1–10.

[17] P. Saleiro, B. Kuester, L. Hinkson, J. London, A. Stevens, A. Anisfeld, K. T. Rodolfa, and R. Ghani, "Aequitas: A bias and fairness audit toolkit," *arXiv preprint arXiv:1811.05577*, 2018.

[18] M. Zehlike, C. Castillo, F. Bonchi, S. Hajian, and M. Megahed, "Fairness measures: Datasets and software for detecting algorithmic discrimination," in *http://fairness-measures.org/*, 2017.

[19] F. Tramer, V. Atlidakis, R. Geambasu, D. Hsu, J.-P. Hubaux, M. Humbert, A. Juels, and H. Lin, "Fairtest: Discovering unwarranted associations in data-driven applications," in *IEEE European Symposium on Security and Privacy*, 2017.

[20] M. Georgopoulos, Y. Panagakis, and M. Pantic, "Investigating bias in deep face analysis: The kanface dataset and empirical study," *Image and Vision Computing*, vol. 102, p. 103954, 2020.

[21] Y. Xu and Z. Jin, "Down-sampling face images and low-resolution face recognition," in *2008 3rd International Conference on Innovative Computing Information and Control*. IEEE, 2008, pp. 392–392.

[22] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *J. Artif. Int. Res.*, vol. 16, no. 1, p. 321–357, Jun. 2002.

[23] K. Ma, X. Wang, X. Yang, M. Zhang, J. M. Girard, and L.-P. Morency, "Elderreact: A multimodal dataset for recognizing emotional response in aging adults," in *International Conference on Multimodal Interaction*, 2019, p. 349–357.

[24] M. Wang, W. Deng, J. Hu, X. Tao, and Y. Huang, "Racial faces in the wild: Reducing racial bias by information maximization adaptation network," in *IEEE Int. Conf. on Computer Vision (ICCV)*, 2019.

[25] M. P. Kim, A. Ghorbani, and J. Zou, "Multiaccuracy: Black-box post-processing for fairness in classification," in *Association for Computing Machinery*, ser. AIES '19, New York, NY, USA, 2019, p. 247–254.

[26] N. Bantilan, "Themis-ml: A fairness-aware machine learning interface for end-to-end discrimination discovery and mitigation," *Journal of Technology in Human Services*, vol. 36, no. 1, pp. 15–30, 2018.

[27] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth, "A comparative study of fairness-enhancing interventions in machine learning," in *Conf. on fairness, accountability, and transparency*, 2019.

[28] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang, "Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," *IBM Journal of R&D*, vol. 63, pp. 4:1–4:15, 2019.

[29] B. W. Schuller, "Multimodal affect databases: Collection, challenges, and chances," in *The Oxford Handbook of Affective Computing*, R. Calvo, S. D'Mello, J. Gratch, and A. Kappas, Eds. Oxford: Oxford University Press, 2015.

[30] H. Gunes and M. Pantic, "Automatic, dimensional and continuous emotion recognition," *Int. J. Synth. Emot.*, vol. 1, no. 1, pp. 68–99, 2010.