Small but Fair! Fairness for Multimodal Human-Human and Robot-Human Mental Wellbeing Coaching

Jiaee Cheong*, Micol Spitale*, Hatice Gunes

***Abstract*—In recent years, the affective computing (AC) and human-robot interaction (HRI) research communities have put fairness at the centre of their research agenda. However, none of the existing work has addressed the problem of machine learning (ML) bias in HRI settings. In addition, many of the current datasets for AC and HRI are 'small', making ML bias and debias analysis challenging. This paper presents the first work to explore ML bias analysis and mitigation of three small multimodal datasets collected within both a human-human and robot-human wellbeing coaching settings. The contributions of this work includes: i) being the first to explore the problem of ML bias within HRI settings; and ii) providing a multimodal analysis evaluated via modelling performance and fairness metrics across both high and low-level features and proposing a simple and effective data augmentation strategy (MixFeat) to debias the small datasets presented within this paper; and iii) conducting extensive experimentation and analyses to reveal ML fairness insights unique to AC and HRI research in order to distill a set of recommendations to aid AC and HRI researchers to be more engaged with fairness-aware ML-based research.**

***Index Terms*—small dataset, fairness, multimodal, well-being coaching, human-robot interaction**

I. Introduction

In recent years, the advancement in machine learning (ML), the availability of large-scale datasets and the enhancement in computing have led to the widespread use of machine-learning prediction systems in our society [1]. However, the problem of bias in machine-learning based tools and systems is becoming an increasing source of concern [2].

There is currently no consensus on how bias should be defined [3, 4]. For instance, "bias" has been used to describe a wide range of concerns, even though each concern is unfair to different groups in different ways and for different reasons. Using gender as an example sensitive attribute, unequal dataset representation, differences in model performance accuracy, or higher prediction probability of prediction of an outcome are often all similarly described as "racial bias".

Scholarly perspective on the definition of bias and fairness varies: some consider bias a subset of fairness (i.e., fairness can exist without group accuracy), whereas others regard fairness as a subset of bias (i.e., fairness necessitates group accuracy). In general, bias is a descriptive term which tells us if there is a distortion or imbalance. On the other hand, fairness is a normative term which tells us whether outcomes align with ethical or legal standards. Bias is generally measured or quantifiable whereas fairness is often subjective or value-laden [3, 5]. Addressing bias is often a step toward achieving fairness, but fairness may require broader ethical consideration beyond just statistical parity [2]. Given the large scope of discussion within this field, it is not possible to thoroughly examine all the nuanced formalisations in this paper.

Readers can refer to the following works for a more in-depth exposition on these topics [5, 6, 2].

The problem of bias and fairness is also becoming an increasingly greater source of concern within both the affective computing (AC) and the human-robot interaction (HRI) research communities [7]. Some of the fairness related concerns highlighted include fairness within a HRI teamwork context [8], robot navigation [9] as well as HRI ethics and robot design [10]. However, this relatively nascent field has yet to consider the fairness-related challenges that occur due to the bias present in ML algorithms deployed within a HRI wellbeing coaching setting.

Within the limited scope of this paper, we only focus on a narrow aspect of fairness. i.e. reducing the group-based probability of positive outcome prediction difference in a model's output, even though ML fairness encompasses considerably more than that. We interpret a reduction in group-based disparity as an improvement in fairness as a difference in group-based fairness measures has the potential to cause harm which does not align with fairness values such as parity and equal opportunity.

As existing bias mitigation approaches chiefly focus on large datasets, they may not be effective for small datasets [4, 2]. However, most of the datasets currently available for AC application scenarios and within HRI contexts are small, i.e., containing just a few hundred instances of data [11, 8]. Figure 1 considers all papers that have been published within the last three editions of the IEEE International Conference on Affective Computing & Intelligent Interaction (ACII) and the ACM/IEEE international conference on Human-Robot Interaction (HRI) respectively. The figure illustrates that papers focusing on small datasets typically represent 40% to 60% of the total papers accepted for presentation at the main conference track. Based on this, we consider any dataset that has less than 40 (median) subjects or 500 (median) samples 'small'. We excluded papers that used large benchmark datasets such as AffectNet. Even if small datasets present challenges for studying fairness, particularly due to the difficulty of distinguishing between bias and sampling error, in such cases, fairness concerns remain relevant and require methodological

*equal contribution, alphabetical order

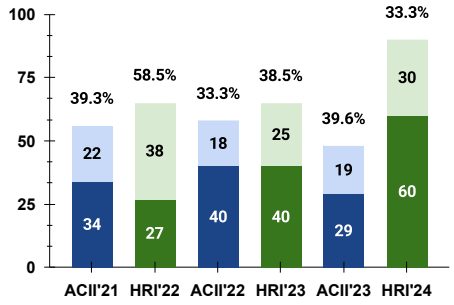

Fig. 1. Proportion of small dataset papers accepted at ACII'21-'23 (in blue) and HRI'22 - '24 (in green). Lighter shade represents the counts for the small datasets whereas the darker shade represents the counts for the bigger datasets.

approaches that account for the constraints of small datasets. Our work explores fairness in this realistic setting, rather than assuming that large datasets will always be available.

Within our ACII 2023 paper [12], we highlighted the ACII community's attempt to be more ethically oriented as exemplified by the mandatory ethics impact statement to guard against the potential risks and harms that could be perpetuated by affect-related technology [1]. We hypothesised and showed that bias exists even for small datasets. Our experiments demonstrated that high-level features were often more informative and more reliable than low-level features and that a multimodal approach is often better than a unimodal approach across both performance and fairness metrics. We also proposed a simple, yet effective method that is able to mitigate against the bias present. In this work, within wellbeing context, we widen the impact of our previous work by extending the experiments towards small datasets in human-robot interaction scenarios.

We do so by introducing the first comprehensive work which explores the problem of bias in a small dataset within a high-stake and sensitive use of wellbeing coaching both within a dyadic human-human and robot-human mental wellbeing coaching setup. We investigate different data augmentation approaches to debias three small temporal multimodal mental wellbeing datasets. We further investigate the contribution of each individual modality (i.e., face, audio, verbal) and the importance of high and low-level features for data-driven applications. The main contributions of this paper are as follows. We conduct the first ML bias and fairness analysis in a HRI context by extending our human-human interaction (HHI) work in ACII'23 [12] to the HRI datasets in this paper. Second,we provide the very first attempt to analyse ML-based wellbeing prediction within the two HRI datasets which are yet to be explored. Third, we provide a thorough multimodal analysis and a feature importance analysis evaluated using both performance and fairness metrics across three different wellbeing coaching datasets. Fourth, we experiment with different data augmentation strategies to reduce the bias in small dataset experimental settings. Lastly, we distil insights and provide guidelines to assist AC and HRI researchers in future small dataset ML studies.

[1]https://acii-conf.net/2022/authors/submission-guidelines/

II. Literature Review

A. Bias and Fairness in Mental Wellbeing

There is only a handful of studies which have looked into bias and fairness in mental wellbeing prediction [13, 14, 15]. As outlined within the introduction, there are many different perspectives on the topic of bias and fairness. Within the context of mental wellbeing literature, similar to the approach that we have adopted in this paper, most of the works have chiefly focused on achieving a fairer outcome via the lens of parity and equal opportunity by reducing the group difference across prediction accuracy and probability of positive class prediction.

Zanna et al. [14] conducted their experiments on data collected in the wild with a specific focus on anxiety prediction. Park et al. [16] analysed bias across gender in mobile mental health assessment and proposed an algorithmic impact remover to mitigate unwanted bias. Bailey and Plumbley [13] attempted to mitigate the gender bias present in the DAIC-WOZ dataset using data re-distribution. [15] examined whether bias exists in existing mental health datasets and algorithms and provided practical suggestions to avoid hampering bias mitigation efforts in ML for mental health. However, all of the existing works consist of relatively large datasets (more than 500 samples or more than 40 subjects) which differ from our small dataset setup. In addition, no investigation has specifically looked into the problem of bias in the context of a human-human and robot-human mental wellbeing coaching.

B. Bias and Fairness in HRI

In contrast with bias and fairness research in mental wellbeing, Most of the existing work on bias and fairness in HRI relate to fairness-related considerations, subject value alignment and normative perspectives within different HRI settings [17, 10, 18]. This is different from the fairness-related challenges that occur due to the disparity in group-level outcome differences produced by the ML prediction model. Londono *et al.* [17] presented a survey on fairness in robot learning. Ogunyale *et al.* [19] conducted experiments to investigate the impact that a robot's "skin colour" can have on human perceptions of the robot's behaviour. Ostrowski *et al.* [10] explored the idea of fairness via the concept of resource allocation. Chang *et al.* [8] investigated fairness within the context of human-robot teaming where fairness is quantified in terms of each member's contribution. Haring *et al.* [20] outlined the implications that a robot's design has on the a human's bias to interact socially with the robots. Alarcon *et al.* [21] investigated the human biases present within human-robot versus human-human trust interactions settings. Lachemaier *et al.* [22] analysed the human automation bias that arose when participants believe a robot's false judgment. Claure *et al.* [23] defined fairness as a constraint on the minimum rate that each human teammate is selected and provided theoretical guarantees on performance. Chang *et al.* [24] showed how participants' perceived fairness is significantly increased when the robot displays effort. Jung *et al.* [25] highlighted the different kinds of insights tasks can yield and how it can be adapted to various human robot collaboration contexts.

However, none of the existing works have investigated the ML accuracy, prediction and outcome disparities within a HRI for wellbeing coaching context which is crucial to achieve fairer outcomes. This study constitutes the first attempt to apply the mitigation of accuracy and prediction disparities on a robot-human wellbeing coaching application.

C. Robotic Mental wellbeing Coaching

Only a small number of studies have explored the use of robotic coaches to support mental wellbeing [26, 27, 28]. Jeong et al. [26] conducted a longitudinal study where Jibo robots provided positive psychology interventions to students in home settings over 7 days. Shi et al. [29] investigated how physical embodiment and personalization affect the perceived quality of text-to-speech (TTS) voices used for mindfulness exercises. Spitale et al. [30] conducted a longitudinal study in which employees of a tech company interact with two different forms of robotic coaches that delivered positive psychology exercises over 4 weeks. Jeong et al. [31] explored how the robot's role (assistant, coach, or companion) affected the therapeutic alliance during wellbeing practice. Axelsson et al. [28] deployed a robotic mindfulness coach at a public cafe, where participants could join robot-led meditation sessions in a group setting. Spitale et al [27] proposed a novel LLM-based multimodal system, namely VITA, that allows robotic coaches to autonomously adapt to the coachee's multi-modal behaviours (facial valence and speech duration) and deliver coaching exercises. This emerging body of research suggests that robotic coaches have the potential to support mental health, but none of the studies have explored the bias of computational models embedded in robotic health coaches.

D. Data Augmentation for Bias Mitigation

Bias can be mitigated at the pre-processing, in-processing or post-processing stage [4]. The proposed method falls under the pre-processing data augmentation category which has proven to be effective in mitigating bias [32]. There is minimal work that focus on mitigating bias for a small dataset setup [33]. For a small dataset problem, [33] leverages on a small annotated dataset to debias a larger dataset. This is distinct from our work as it focuses specifically on an item recommendation system. Existing research has indicated that re-sampling outperforms reweighting for correcting sampling bias [34]. Given the above, we propose a simple re-sampling or data augmentation method based on the mixup method proposed in [35]. *Mixup* has proven to be a simple yet highly effective method to address challenges ranging from robustness [36], fairness [37] and regularisation [38]. As a result, *Mixup* has been frequently used as a benchmark for new data augmentation techniques and there are recent works proposing new variations of the original method [37].

III. Problem Formulation

We adopt a machine learning approach, where the goal is to predict a correct outcome $y_i \in Y$ from input $\mathbf{x}_i \in X$ based on the available dataset D for individual $i \in I$. In our setup, $y_i \in Y$ is thus the outcome where $Y = 1$ denotes "high-PA" (i.e., high positive affect, indicative of higher levels of mental wellbeing) whereas $Y = 0$ simply denotes otherwise. The fairness measure of a model M is then evaluated according to the subgroups of individuals defined by their sensitive attributes A gender and race in this work. In our experiments, both sensitive attributes analysed are binary. They belong to the majority group, e.g.: $A_{race} = 1$ if they are White or $A_{race} = 0$ if otherwise. $\hat{Y}$ denotes the predicted class. However, we wish to caution that this only constitutes a small aspect of fairness. Section VIII provides a more nuanced discussion of its interpretation.

A. Fairness Measures

As highlighted within the introduction, there are many definitions and formalisation for bias and fairness. Within the context of this work, we adopt the setting in which bias is the disparity that we wish to address to achieve fairer systems which are measured using commonly used fairness measures comparable with existing works [39, 14, 12]. We use $A = 0$ to denote the minority and $A = 1$ to denote the majority group throughout the paper.

- **Equal Accuracy** (EA), a group-based metric, is used to compare the group fairness between the models. This can be understood as the accuracy gap between the majority and the minority group:

$$EA = |MAE(\hat{Y}|A = 1) - MAE(\hat{Y}|A = 0)|, \quad (1)$$

  where MAE represents the Mean Absolute Error (MAE) of the classification task of each sensitive group.
- **Disparate Impact** (DI), measures the ratio of positive outcome ($\hat{Y} = 1$) for both the majority and minority group as represented by the following equation:

$$DI = \frac{Pr(\hat{Y} = 1|A = 0)}{Pr(\hat{Y} = 1|A = 1)} \quad (2)$$

In addition to being the most commonly used metrics, another key motivation for selecting these two fairness measures is that they represent different aspects of fairness. EA evaluates fairness based on the model's predictive performance measured in terms of difference in accuracy across the different subgroups, whereas DI evaluates fairness based on the predicted outcomes $\hat{Y}$ of each subgroup. In general, the closer EA is to 0.00, the fairer the results. The closer DI is to 1.00, the fairer the results. However, a more nuanced interpretation of these measures is required. For instance, one important assumption missing from the DI definition is that it assumes there is an equal ratio of positive outcomes across groups. This comes down to differences in the base rates across groups in the population. In our experiments, the base rates are similar. If the base rates are different, DI may not be the most appropriate fairness measure to evaluate fairness. We will address such nuances in greater detail within Section VIII of the paper.

B. Proposed Method: MixFeat

Our proposed methodology (MixFeat) is based on the data augmentation technique proposed by [35]. Given a dataset of

size N where A represents the audio cue, F represents the facial cue and V represents the verbal cue, the new training sample (A_k, F_k, V_k) is therefore generated as follows:

$$\begin{aligned} A_k &= \lambda_A \cdot A_i + (1-\lambda_A) \cdot A_j \\ F_k &= \lambda_F \cdot F_i + (1-\lambda_F) \cdot F_j \\ V_k &= \lambda_V \cdot V_i + (1-\lambda_V) \cdot V_j \end{aligned} \tag{3}$$

where $i, j \in \{1, ...N\}$, $i \neq j$ and $\lambda_A, \lambda_F, \lambda_V \sim$ Beta(0,1) as suggested by the original paper as suggested by the authors of the original paper, Zhang *et al.* [35]. It is worth noting that both the α and β parameters in Beta(α, β) distribution must be bigger than 0 [35]. It is worth nothing that both the α and β parameters in Beta(α, β) distribution must be bigger than 0. We use the above method to generate synthetic samples for the minority group to obtain balanced samples across the sensitive attributes of race and gender.

Although *MixFeat* bears some similarity to other methods such as the Synthetic Minority Over-sampling Technique (SMOTE) [40], the key distinction is that in SMOTE, synthetic samples were generated by taking the difference between the feature vector of the sample under consideration and its nearest neighbour and then multiplying this difference by a random number between 0 and 1 before adding it back to the feature vector under consideration. However, popular methods such as SMOTE may lead to over-generalization with high variance [41] and are largely ineffective in dealing with highly imbalanced datasets [42]. 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 [43]. This motivates us to propose a new method to address this challenge given our small dataset setting which may contain too few instances of the minority group for SMOTE to work as intended.

On the other hand, *MixFeat* operates by interpolating features from the same sensitive attribute group. The intuition behind this method is that if we generate new samples by mixing up features from other samples with the same sensitive attribute, the new samples will inherit the sensitive-attribute specific features. This bears conceptual similarity to Fair-SMOTE [44] which restricts the selection of nearest neighbour to the same group or sensitive attribute rather than a heterogeneous pool such as SMOTE [40]. However, the difference is that Fair-SMOTE uses two hyperparameters "mutation amount" and "crossover frequency" of *fixed values* to generate synthetic samples which may require slightly more assumptions to be made during the hyper-parameter tuning process. On the other hand, MixFeat uses a Beta distribution sampling method for selecting the mixture parameter. This method preserves the relation between the synthetic samples and supervision signal which gives the algorithm more samples to learn from without imposing strong assumptions [35]. Moreover, Fair-SMOTE aims to "extrapolate *all* the variables by the same amount" [44] whereas MixFeat aims to only interpolate variables *within the same modality*. Figure 2 outlines the experimental setup and how the method is integrated into the overall classification pipeline.

IV. Datasets and Methods

Given the small dataset sizes, we limit our problem to a binary classification problem with details described below.

A. Datasets

1) The AFAR BSFT Dataset: We collected a dataset of human-human dyadic interactions between a human wellbeing coach and 11 participants over four weeks. The human wellbeing coach was instructed to deliver a Brief-Solution Focused Therapy (BSFT) style coaching, asking participants to focus on solutions rather than analysing the problem [45] for about 20 minutes. After each session, we asked participants to complete the Positive And Negative Affect Scale (PANAS) [46] to evaluate their positive and negative affect.

a) Data Collection: 11 participants were recruited via email advertising of the University of Cambridge. We conducted the study in a dedicated room where a human wellbeing coach and one participant were seated in front of each other. Video recordings were done using two external cameras, one facing the participant and the other facing the human coach that can be used for further analysis (beyond the scope of this paper) on dyadic interactions during the coaching practice. We collected 44 videos (11 participants $\times$ 4 weeks, 20 mins per session) of dyadic wellbeing coaching interactions. 3 out of 44 sessions were excluded due to technical issues (e.g., corrupted video or audio recordings).

b) Sensitive Groups: Two human annotators labelled the gender and race of the participants (with a 100% agreement). This resulted in 7 participants being labelled as men and 4 as women, and 8 participants being labelled as Whites, and 3 as non-Whites. Note that we relied on external annotations for gender and race, which may have introduced labeling bias, as these attributes are not directly observable [47]. Additionally, since annotators labeled only binary gender, this may have introduced further bias.

2) The AFAR Robocoaching 2022 Dataset (AFAR-RC22): In 2022, we collected a dataset of human-robot interactions between a robotic mental wellbeing coach and 26 participants over four weeks in a tech company (Cambridge Consultants Inc.) [30]. The robotic wellbeing coach delivered four positive psychology exercises once a week – savouring, gratitude, accomplishments, and optimism about the future – that lasted around 10 minutes each. As we did for the AFAR BSFT dataset, among other measures, we asked participants to fill out the PANAS questionnaire after each interaction session with the robotic coach. The robotic coach was pre-programmed to conduct the positive psychology practice following predefined steps regardless of the employees speech. For example, when the robotic coach asked the employee to share what they have been grateful for during the last week, the robot asked the same follow up question to all employees without adapting the coaching to what has been said by the employees.

a) Data Collection: Cambridge Consultants Inc advertised the study via their communication channels and the participation was on voluntary-basis. 26 participants that took part were healthy employees of the company. Please refer to our paper [30] for more information about this study and

the recruitment and screening process. Employees interacted once a week with a robotic coach that delivered positive psychology exercises over four weeks. This was a between-subject study in which employees were randomly assigned to interact either with a QT robot (humanoid-like appearance) or a Misty II robot (toy-like appearance). Video recordings were done using an external camera that captured the employees' behaviours during the robotic wellbeing coaching [48]. We collected a total of 104 videos (26 participants×4 weeks, 10 mins each) of robotic wellbeing coaching sessions. 4 of them were excluded due to technical issues (e.g., corrupted audio-visual recordings).

b) Sensitive Groups: Participants self-reported their gender before the study: 6 participants self-reported as woman, 1 as non-binary person, and 19 as men. While for race, two human annotators labelled the race of participants (with again 100% agreement), that resulted in 21 participants labeled as Whites, and 5 participants as non-Whites.

3) The AFAR Robcoaching 2023 Dataset (AFAR-RC23): We collected the second dataset on longitudinal robot coaching in 2023 as reported in [49, 27] in the same tech company. This dataset collated data recorded into two study. The first study reported in [49] involved 12 participants – who had already interacted with a robotic coach in [30]. The second study involved 17 new participants who had never interacted with a robotic coach before.

a) Data Collection: A total of 29 participants were involved in the two studies with the same recruitment process reported in Section IV-A2. Please refer to the following papers for more details about the two studies [49, 27]. Employees interacted once a week with the QT robot (the robotic platform chosen for these two studies, see [27]) that delivered four positive psychology exercises over four weeks – savouring, gratitude, accomplishments, and one door closes one door opens. Audio-visual data were collected via an external camera that captured the face and body of the employees interacting the the robotic wellbeing coach. We collected a total of 116 videos (29 participants x 4 week, 10 mins for each session). 1 of them was excluded due to technical issues.

b) Sensitive Groups: Participants self-reported their gender: (study 1) 3 women, 1 non-binary, and 8 men; and (study 2) 7 women, and 10 men. While for the race, two human annotators have labelled the dataset with a full agreement as follows: (study 1) 3 non-Whites and 9 Whites; (study 2) 1 non-White and 16 Whites.

4) Annotations: We assessed the participants' positive affect using the self-report results of the PANAS questionnaire [46] for all three datasets, which has been widely used by practitioners to identify strengths and concerns in mental wellbeing. We computed the positive affect (PA) and negative affect (NA) sub-scales according to the manual in [46]. We set the threshold value to 33.3, corresponding to the mean value for the American population [46], and we then classified the videos collected into "high-PA" and "low-PA". This resulted in: (1) AFAR-BSFT DB: 17 videos for the "low-PA" and 26 videos for the "high-PA" class; (2) AFAR-RC22 DB: 45 videos for the "low-PA" and 57 videos for the "high-PA" class; and (3) AFAR-RC23 DB: 51 videos for the "low-PA" and 64 videos for the "high-PA" class.

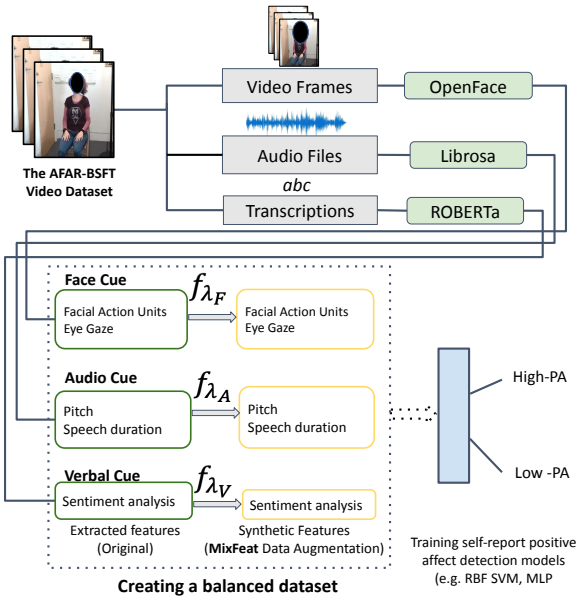

Fig. 2. The model pipeline with our proposed data augmentation technique: **MixFeat**. After extracting the high-level features from the dataset, we generate synthetic sample features using Equation 3. Each modality's feature generation process is chiefly governed by their respective $\lambda \sim$ Beta(0,1) parameters. Examples of high-level features include Facial Action Unit (FAU) and gaze for the face modality, pitch and speech duration for the audio modality and sentiment for the textual modality. A more comprehensive overview can be found in Table III.

B. Self-report Affect Detection

1) Dataset Pre-processing: Before extracting the features, we split the audio and video recordings of the three datasets. We asked a human annotator to transcribe the dyadic interactions between the human coach and the participants manually for the AFAR-BSFT DB. The annotator also took note of the timestamp of the speech so that we were able to diarize the audio files. For the AFAR-RC22 dataset, we did not transcribe coachee interactions during robotic coaching. However, for the AFAR-RC23 dataset, we automatically transcribed the coachees' dialogue during interactions with the robotic coach, enabling adaptive conversation responses.

2) Multi-modal Feature Extraction: We extracted the facial features using OpenFace 2.0 [50] – which represents one of the state-of-the-art tools for extracting facial features within the affective computing community, e.g., in a recent work [51] – resulting in the following: eye gaze directions, the intensity and presence of 17 facial action units (FAUs), facial landmarks, head pose coordinates, and point-distribution model (PDM) parameters for facial location, scale, rotation and deformation, resulting in 709 facial features. We adopted this facial features extraction strategy for all three datasets. We used librosa[2] to extract the audio features for the AFAR-BSFT dataset, namely pitch, speech duration, 128 Mel spectrograms, 20 MFCC, 20 delta MFCC, spectral centroid, and RMS,

[2]https://librosa.org/doc/latest/index.html

which results in 172 audio features, as in previous works, e.g., [52]. While for AFAR-RC22 and AFAR-RC23 DBs, we used openSMILE[3] to extract audio features using the GeMaPs method, that includes e.g., loudness, alpha ratio, hammarberg index, slop, spectral flux and MFCC, that resulted in a total of 25 features, because we have already extracted such features for our previous works [30, 27]. We used ROBERTa[4] to extract the predicted sentiment for all three datasets from the participants' transcriptions resulting in 2 verbal features (label and probability), as in [11].

3) Pre-processing: We first removed constant and null features to prepare the multi-modal features for the machine learning models. Then, we decided to condense the temporal information of each video clip into statistical descriptors as in [11], computing a fixed-length vector for each multi-modal feature of each clip that consists of mean, median, standard deviation, minimum, maximum, and auto-correlation with 1-second lag, resulting in: (1) AFAR-BSFT DB: a facial feature vector with size $41 \times 709 \times 6$, in an audio feature vector with size $41 \times 172 \times 6$, and in a verbal feature vector with size $41 \times 2 \times 6$; (2) AFAR-RC22 DB: a facial feature vector with size $100 \times 709 \times 6$, in an audio feature vector with size $100 \times 75 \times 6$, and in a verbal feature vector with size $100 \times 2 \times 6$; and (3) AFAR-RC23 DB: a facial feature vector with size $115 \times 709 \times 6$, in an audio feature vector with size $115 \times 75 \times 6$, and in a verbal feature vector with size $115 \times 2 \times 6$.

TABLE I
UNI- (TOP) AND MULTI-MODAL (BOTTOM) HIGH VS LOW-LEVEL FEATURE MODELING RESULTS FOR THE **AFAR-RC22 DATASET**. VALUES IN BOLD DENOTE THE BEST OUTCOME ACROSS THE EXPERIMENTS. * DENOTES p-VALUES OF ≤ 0.05

	Uni-modal			
	Face		Audio	
	Low	High	Low	High
R-Acc	0.56	**0.62***	0.52	0.52
R- F1	0.68	**0.69**	**0.65**	0.63
M-Acc	0.55	**0.65***	0.57	**0.58**
M- F1	0.62	**0.68**	**0.65**	0.64

	Face and Audio					
	Early		Soft Voting		Stacking	
	Low	High	Low	High	Low	High
R-Acc	**0.58**	0.55	0.47	**0.53***	0.52	**0.54**
R-F1	0.66	**0.71**	0.52	**0.61***	0.59	**0.62**
M-Acc	**0.59***	0.52	0.55	0.55	**0.59**	0.55
M-F1	**0.65***	0.56	0.60	**0.71***	0.60	**0.71***

4) Feature Selection: We defined the high-level and low-level features as interpretable (e.g., facial action unit, pitch) and not-interpretable (e.g., spectral features) to select the most informative ones for the positive affect detection model [53]. The low-level vs high-level features are summarised in Table III. Given the differences in dimensionality between low-level and high-level features, we conducted a principal component analysis (PCA) to reduce the size of the features while keeping 80% of the information. The PCA analysis resulted in:

[3]https://www.audeering.com/research/opensmile/
[4]https://huggingface.co/docs/transformers/model_doc/roberta

TABLE II
UNI- (TOP) AND MULTI-MODAL (BOTTOM) HIGH VS LOW-LEVEL FEATURE MODELING RESULTS FOR THE **AFAR-RC23 DATASET**. R: RBF SVM. M: MLP. VALUES IN BOLD DENOTE THE BEST OUTCOME ACROSS THE EXPERIMENTS. * DENOTES p-VALUES OF ≤ 0.05

	Uni-modal					
	Face		Audio		Verbal	
	Low	High	Low	High	Low	High
R-Acc	0.63	**0.70***	**0.55**	0.54	**0.57**	0.56
R- F1	0.74	**0.78**	**0.66**	0.65	0.68	0.68
M-Acc	0.66	**0.73***	0.51	**0.53**	0.60	0.60
M- F1	0.75	**0.78**	0.61	**0.69***	0.69	0.69

	Face and Audio					
	Early		Soft Voting		Stacking	
	Low	High	Low	High	Low	High
R-Acc	0.59	**0.60**	0.65	**0.72***	0.69	**0.72**
R-F1	0.74	**0.75**	0.77	**0.78**	0.77	**0.78**
M-Acc	**0.59***	0.50	0.63	**0.73***	0.62	**0.73***
M-F1	**0.70***	0.62	0.74	**0.80***	0.73	**0.79***

- **AFAR-BSFT.** i) 5 principal components (PCs) for high-level features for face, 10 PCs for low-level features for face, and ii) 2 features for high-level features for audio (no PCA conducted because the number of high-level audio features was already small, i.e., equal to 2), and 3 PCs for low-level features for audio.
- **AFAR-RC22 DB:** i) 19 PCs for high-level features for face, 5 PCs for low-level features for face, and ii) 4 PCs for high-level features for audio, and 3 PCs for low-level features for audio.
- **AFAR-RC23):** i) 17 PCs for high-level features for face, 6 PCs for low-level features for face, and ii) 3 PCs for high-level features for audio, and 6 PCs for low-level features for audio.

5) Data Fusion Strategies: We explored different state-of-the-art data fusion strategies [54] for all three datasets. We experimented with early fusion, which consisted of concatenating features from different modalities that resulted in a single vector of features, and different late fusion strategies, namely majority voting (soft and hard) and stacking (soft and hard). In majority voting, the final decision is made according to the most frequent class label predicted across the different uni-modal models (hard) or the classifier whose predicted class probability is the highest across the different uni-modal models (soft). In stacking, the final decision is made by another classifier (e.g., logistic regression model) fed by either the predicted class label (hard) or the predicted class probabilities (soft) of each uni-modal model.

V. MODELING AND BIAS ANALYSIS RESULTS

A. Modeling and Feature Selection

We first conducted experiments using various ML techniques as in [11] – namely logistic regression, linear support vector machine (SVM), random forest tree, bagging, XGBoost, AdaBoost, decision tree, radial basis function support vector machine (RBF-SVM), multi-layer perceptron (MLP), and long-short term memory (LSTM) neural network – and

TABLE III
OVERVIEW OF THE DIFFERENT HIGH VS LOW-LEVEL FEATURES USED.

Dataset	Low-level Features	High-level Features
Face	Facial landmarks, head pose coordinates, and point-distribution model (PDM) parameters.	Facial action units and gaze.
Audio	128 Mel spectrograms, 20 MFCC, 20 delta MFCC, spectral centroid, and RMS for AFAR-BSFT dataset and MFCC, shimmer, jitter, F0, F1, F2, and F3 as audio features for the AFAR-RC22 and 23 DBs	Pitch and speech duration for AFAR-BSFT dataset and loudness, alpha ratio, hammarberg index, slop, spectral flux for the AFAR-Robo Coaching 2022 and 2023 DBs.
Verbal		Sentiment of the speech

TABLE IV
UNIMODAL AND MULTIMODAL DEBIASING RESULTS FOR THE **AFAR-RC22 DATASET**. ABBREVIATIONS. R: RBF SVM. M:MLP. UAR: UNWEIGHTED AVERAGE RECALL. VALUES IN BOLD DENOTE THE BEST OUTCOME ACROSS THE THREE SETS OF EXPERIMENTS.

	Original				Baseline Comparison				Proposed Method			
	Face		Audio		Face		Audio		Face		Audio	
	R	M	R	M	R	M	R	M	R	M	R	M
Overall Acc	0.39	0.54	0.48	0.52	0.56	0.51	0.54	0.56	**0.59**	0.55	**0.59**	**0.59**
Overall F1	0.55	0.60	0.58	0.59	0.66	0.57	0.58	0.51	**0.68**	0.65	**0.68**	0.67
Overall UAR	0.57	0.61	0.59	0.60	0.69	0.58	0.59	0.52	**0.70**	0.67	**0.70**	0.68
EA_{Gender}	**0.01**	0.10	0.13	0.12	0.08	0.06	0.08	0.08	0.02	0.09	0.18	0.13
EA_{Race}	0.04	0.05	**0.01**	0.07	0.22	0.21	0.17	0.17	0.17	0.23	0.26	0.19
DI_{Gender}	1.02	1.11	**1.00**	0.99	1.08	0.75	0.82	0.73	1.01	0.77	0.93	0.78
DI_{Race}	0.83	0.84	1.09	1.45	0.91	1.13	1.11	1.24	0.97	1.10	**0.99**	1.20

	Early		Soft Voting		Stacking		Early		Soft Voting		Stacking		Early		Soft Voting		Stacking	
	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M
Overall Acc	0.54	0.50	0.62	0.57	0.55	0.60	0.61	0.53	0.59	0.56	0.63	0.63	**0.66**	0.54	0.60	0.61	**0.68**	0.61
Overall F1	0.69	0.55	0.64	0.63	0.62	0.57	0.70	0.64	0.68	0.65	0.68	0.71	**0.80**	0.69	0.83	**0.84**	0.77	**0.78**
Overall UAR	0.75	0.55	0.64	0.63	0.62	0.59	0.73	0.66	0.70	0.67	0.69	0.74	**0.78**	0.69	0.83	**0.84**	0.77	**0.82**
EA_{Gender}	0.08	0.07	0.08	0.09	**0.00**	0.17	0.08	0.02	0.11	0.08	0.14	0.02	**0.01**	0.03	0.13	**0.01**	0.25	0.06
EA_{Race}	0.24	0.10	**0.05**	0.28	0.04	0.11	0.22	0.17	0.19	0.17	0.11	0.05	0.13	**0.03**	0.17	**0.05**	0.17	**0.01**
DI_{Gender}	0.93	1.21	**1.04**	0.83	1.34	1.37	**0.94**	0.71	0.85	0.73	**0.98**	0.87	0.93	0.85	0.88	0.77	1.18	0.83
DI_{Race}	1.07	0.90	**1.03**	1.37	0.81	1.09	1.04	1.19	1.07	1.24	1.09	1.21	1.09	**1.04**	1.11	1.26	**1.01**	1.27

validating them with two different cross-validation approaches (i.e., 5-fold CV and leave-one-subject-out (LOSO)). Our results showed that the best models were RBF-SVM and MLP among the machine learning techniques we experimented with. Due to space constraints, we only report the best performing model results and analyses in the following sections.

B. Low vs High Level Feature Analysis

1) AFAR-BSFT: We trained different experimental models with either the high or low-level features, and compared their performances. All comparisons were conducted using the t-test at a 5% (i.e. 0.05) significance level. The top part of Table I of the Supplementary Material reports the results of the uni-modal models, while the bottom part of Table I of the Supplementary Material reports the results of the multi-modal (i.e., face and audio) models. The results can be found in the Supplementary Material and the detailed analysis can be found in our previous paper [12].

2) AFAR-RC22: For the AFAR-RC22 dataset, with reference to Table I (top), we see a similar trend where the high-level features are better for the face modality. Across the audio modality, though the low features seem to perform better for F1 scores across both the RBF SVM and MLP methods, the gap in results are minor. Across the multimodal setup, with reference to Table I (bottom), we see a similar trend with the AFAR-BSFT dataset. For the early fusion strategy, low-level features seem to perform better whereas for the late fusion strategies (soft voting and stacking) high-level features performed better.

3) AFAR-RC23: For the AFAR-RC23 dataset, with reference to Table II (top), we see a similar trend where the high-level features are better for the face modality. Across the audio modality, RBF SVM performed better with low level features and MLP performed better for high level features. Across the verbal modality, both models seem to produce similar results across both high and low level features. Across the multimodal setup, with reference to Table II (bottom), we see a similar trend with the two other datasets. The key difference is that across early fusion, the RBF SVM performed better using high level features whereas the MLP performed better with low level features. For the late fusion strategies (soft voting and stacking), high-level features performed better just as before.

C. Uni-modal vs Multi-modal Analysis

We conducted several experiments to compare uni-modal and multi-modal (with either early or late fusion) approaches.

1) AFAR-BSFT: The results can be observed in Table II of the Supplementary Material, and our previous paper [12].

TABLE V
UNIMODAL AND MULTIMODAL DEBIASING RESULTS FOR THE **AFAR-RC23 DATASET**. ABBREVIATIONS. R: RBF SVM. M:MLP. UAR: UNWEIGHTED AVERAGE RECALL. VALUES IN BOLD DENOTE BEST OUTCOME ACROSS THE THREE SETS OF EXPERIMENTS.

	Original						Baseline Comparison						Proposed Method					
	Face		Audio		Verbal		Face		Audio		Verbal		Face		Audio		Verbal	
	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M
Overall Acc	0.46	0.55	0.48	0.46	0.53	0.55	0.57	0.49	0.54	0.57	0.54	0.57	**0.63**	0.57	0.57	**0.59**	0.59	**0.63**
Overall F1	0.52	0.49	0.49	0.45	0.53	0.60	0.66	0.55	0.54	0.49	0.54	0.49	**0.71**	0.66	0.64	**0.66**	**0.66**	0.41
Overall UAR	0.55	0.52	0.50	0.47	0.55	0.62	0.70	0.56	0.55	0.49	0.55	0.49	**0.74**	0.68	0.65	**0.67**	**0.68**	0.42
EA_{Gender}	0.16	0.21	0.21	0.13	0.04	0.18	0.11	**0.02**	0.06	0.04	0.06	0.04	0.23	0.19	**0.01**	0.04	0.07	**0.03**
EA_{Race}	0.05	**0.04**	0.07	**0.01**	0.07	0.03	0.25	0.17	0.15	0.18	0.15	0.18	0.37	0.27	0.10	0.13	0.05	0.06
DI_{Gender}	1.34	1.41	1.24	1.03	1.22	0.96	1.27	0.79	1.08	0.76	**1.08**	0.76	**1.14**	0.84	**1.00**	0.75	1.22	0.84
DI_{Race}	**0.98**	0.90	1.17	1.34	1.31	1.10	0.78	1.07	0.85	**1.10**	0.85	**1.10**	0.86	1.08	**0.91**	1.20	0.88	1.17
	Early		Soft Voting		Stacking		Early		Soft Voting		Stacking		Early		Soft Voting		Stacking	
	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M	R	M
Overall Acc	0.49	0.42	0.60	0.55	0.57	0.61	0.63	0.54	0.61	0.59	0.62	0.60	**0.67**	0.62	0.62	**0.70**	**0.68**	0.59
Overall F1	0.61	0.41	0.52	0.61	0.55	0.57	0.70	0.63	0.68	0.66	0.66	0.68	**0.73**	0.64	0.68	**0.73**	**0.70**	0.67
Overall UAR	0.63	0.41	0.52	0.61	0.57	0.60	0.73	0.66	0.70	0.68	0.67	**0.70**	**0.76**	0.64	0.71	**0.74**	**0.70**	**0.70**
EA_{Gender}	0.05	0.07	0.07	**0.05**	**0.02**	0.16	0.02	0.05	**0.05**	0.07	0.04	0.04	**0.00**	0.02	0.06	**0.05**	0.03	**0.02**
EA_{Race}	**0.02**	0.12	0.02	0.31	0.03	0.10	0.14	0.12	0.12	0.14	0.06	**0.03**	0.10	0.07	**0.03**	0.07	0.07	0.05
DI_{Gender}	1.06	1.52	**1.08**	0.75	1.48	1.51	0.88	0.66	0.79	0.74	0.94	0.87	**1.01**	0.71	0.89	0.90	1.04	**0.97**
DI_{Race}	0.78	0.50	1.09	1.37	0.82	1.10	**1.13**	1.24	**1.07**	1.25	1.06	1.23	1.13	1.21	1.16	1.14	**1.04**	1.16

2) AFAR-RC22: We observe from Table IV that overall, a multimodal approach can lead to better results than a unimodal approach across both performance and fairness. For instance, all of the original multimodal fusion methods (early fusion, soft voting and stacking) gave improvements across accuracy and unweighted average recall (UAR) compared to the best performing unimodal modality and model. Across fairness, stacking seems to produce the best improvement across the EA metrics whereas soft voting seems to produce the best improvement across the DI metrics.

3) AFAR-RC23: We see that previous observations are generally consistent. We see from Table II that a multimodal approach is often better than the best performing unimodal approach across most measures. The best performing accuracy results improved from 0.57 to 0.61 and the best performing UAR results improved from 0.62 to 0.63. Across fairness, the best performing original unimodal EA_{Gender} results improved from 0.04 to 0.02.

VI. DEBIASING APPROACH AND RESULTS

To provide a comparison to our proposed method, we use a baseline data balancing method to mitigate the bias present. We employ a similar data balancing method as [39]. We re-sample the minority group by randomly oversampling data-points to obtain an augmented dataset with samples balanced across both sensitive attributes. The implementation of our proposed method is similar to that of the baseline method. The key difference is that instead of randomly oversampling data points, we generate synthetic samples according to the method outline in Equation 3. We implemented this data balancing and proposed method similar to all three DBs.

A. AFAR-BSFT

After data balancing, we retrain the models and capture the results in Table II of the Supplementary Material. Looking at Table II (top) of the Supplementary Material, across the unimodal experiments, we see our method consistently produces a more accurate and fairer outcome across most metrics for both sensitive attributes compared to the baseline. Within the multimodal approach depicted in Table II of the Supplementary Material (bottom), we see that this gap in predictive and fairness performance is diminished.

B. AFAR-RC22

Within a unimodal setting, we see from Table IV (top) that although our proposed method does not always give the best performance and predictive fairness, overall, it gives frequent improved performance and predictive fairness compared to the other methods. This is also true within a multimodal setting. Across fairness, our results are also better compared to baseline. For instance, across early fusion, our proposed data augmentation method resulted in the fairest EA_{Gender} score of 0.01 with the RBF SVM classifier and the fairest EA_{Race} score of 0.03 with the MLP classifier. Within the multi-modal approach in Table IV (bottom), we see that our proposed method produces improved results across most performance metrics. Across fairness, though our results perform better across most fusion methods we see that the baseline methods produces better fairness scores across DI_{Gender} across both the early fusion (0.94) and stacking (0.98) approaches.

C. AFAR-RC23

We see a consistent trend across the unimodal results in Table V (top). For instance, for the face modality, we achieved an overall accuracy, F1 and UAR of 0.63, 0.71 and 0.74 respectively. Across the audio modality, our proposed method also produced the best results across EA_{Gender}, DI_{Gender} and DI_{Race} with a score of 0.01, 1.00 and 0.91 respectively. Across the multimodal results, V (bottom), our proposed

TABLE VI
RECOMMENDATIONS FOR AFFECTIVE AND HRI RESEARCHERS FOR REDUCING ML BIAS IN SMALL DATASETS.

Research Field	Recommendations	Why?	How?
ML Training	(**R1**) Train human-centric models on a balanced dataset	The problem of bias is often associated with data imbalances [44].	Ensure that participants are balanced across sensitive groups (e.g., gender, race) or perform data balancing methods as needed (e.g., data augmentation [12]).
ML Features Selection	(**R2**) Use higher level features making use of multimodality	The high-level features often include less noisy information and make easier for the model to learn the representations in small datasets if more information (i.e., multimodal) is used.	Extract and train the model using multimodal high-level features by experimenting different fusion strategies [55]
ML Modeling	(**R3**) Employ a variety of ML models and evaluation and fairness metrics	Past works [55] showed that specific models work better on certain modalities across specific metrics.	Conduct experiments using a variety of different models and metrics [11].
AC/Robotic System Design	(**R4**) Balance the trade off between AC/robotic system complexity and fairness	Adaptive and more complex AI-based system may be more difficult to debias [56].	Adaptive models that need to be embedded in a robot should be tested for their fairness in advance [57]. When this is not applicable, on-the-fly model should have embedded bias mitigation strategies.
AC/Robotic System Design & Ethics	(**R5**) Define field and context specific ethical principles when designing / deploying AC systems for wellbeing	The ethical research in robotics studies the consequences of deploying robots in social contexts and of interacting with humans. The study and definition of ethical guidelines can help building fair HRI [17]	Adopt human-centric approach, like value-sensitive design [58], to distill main ethical recommendations to use robots in a specific context [59]

method also produces good results compared to the baseline data augmentation method With reference to Table V (top), we see that both data augmentation methods were effective at improving performance and reducing bias. Within the multi-modal approach depicted in Table V (bottom), we see a similar trend. In general, our proposed method performs better than the baseline across most fusion strategies. We see from our results that the complexity of optimising for both performance and fairness is exacerbated within a small dataset setting.

VII. RECOMMENDATIONS AND CONCLUSION

In addition to our findings reported in [12], we have also noted that as much as it is possible to debias small datasets, the performance of a model, as measured using the standard metrics such as accuracy an F1, is still very correlated and dependent on the size of the dataset. AFAR-RC22 (26 participants with a total of 101 datapoints) and AFAR-RC23 (29 participants with a total of 116 datapoints) are bigger than AFAR-BSFT (11 participants with a total of 41 datapoints). We see from our results that the same models trained on the AFAR-RC22 and AFAR-RC23 datasets perform better across measures such as overall accuracy and overall F1. In addition, the effects of data augmentation and data balancing in these dataset produced greater improvements in performance metrics compared to the BSFT dataset. This is likely due to the fact that we had to generate more synthetic samples in order to balance the samples across the different sensitive attribute groups for the larger datasets (AFAR-RC22 and AFAR-RC23). From our findings, we distilled a set of recommendations **(R1 - R5)** in Table VI that can be used by AC and HRI researchers to integrate and address fairness-related concerns within their ML-based research when working with small datasets.

Note that these recommendations are specific for HRI contexts that uniquely differ from other human-centred computing fields. HRI involves a physical agent (i.e., a robot) that shares the same environment as humans, and as such influences how bias and fairness manifest with respect to more limited interactions like virtual or screen-based, typical of traditional AC. In HRI, the robot's physical presence and its ability to engage through multiple channels (like speech, gestures, and movement) create a more complex and dynamic interaction space, where issues of bias and fairness extend beyond what is seen in purely virtual systems.

R1: Train human-centric models on a balanced dataset when small. Our results suggest that employing an imbalanced human dataset may lead to fairness issues when training ML models. Small datasets are commonly utilised in studies within Affective Computing (AC) and Human-Robot Interaction (HRI), particularly in contexts related to wellbeing [27]. This is also supported by literature that highlighted how the problem of bias is often associated with data imbalances [44]. Curating the dataset by ensuring balanced participation and representation across sensitive groups (e.g., gender, race) during data collection is crucial for fairness. When this is not applicable, data balancing techniques can be employed. Simple methods such as data augmentation via upsampling are capable of significantly improving the lack of fairness present.

R2: Make use of multimodal and higher level features. Our results show that fairness has been improved by using multi-modal and high-level features. Small AC and HRI datasets often include audio-visual recordings and sometimes physiological signals (e.g., EEG, heart rate variability, etc.). As such, they encompass different modalities that can be combined to provide more information for a machine learning algorithm to learn from. Past works [60, 55] have shown how multi-modal models performed better than uni-modal ML models. Past work that has considered a different definition of fairness (e.g., minimising bias) showed that in some context uni-modal features may work better because multi-modal features tend to cross-contaminate models with biases from each

modality [61]. Additionally, high-level interpretable features are beneficial for small datasets because they allow human stakeholders to verify the ML model's results.

R3: Employ a variety of ML models and evaluation and fairness metrics. Our results highlight the importance of exploring a variety of machine learning models and evaluating them using diverse evaluation and fairness metrics. A model that demonstrates strong fairness across diverse subgroups may exhibit lower accuracy compared to a model optimized solely for high performance [62]. It is crucial for each researcher to determine the appropriate bias-accuracy trade off [63] that best suits the specific task and context at hand [47]. Specific ML models may provide more favourable results across certain performance and fairness metrics. Beyond data augmentation, future work can also investigate other ML bias mitigation approaches such as multi-task learning using both performance and fairness [63, 62], or via active learning from human feedback or correction [64]. It will be ideal to experiment with different strategies as different strategies may be suitable for different tasks based on the different AC / HRI context.

R4: Balance the trade off between AC/robotic system complexity and fairness. Our results show that the adaptive capability of robotic systems (used in AFAR-RC23) may have impacted the debiasing strategies in different ways. AFAR-RC22 and AFAR-RC23 shared the same study design; the only difference was in the robotic system's capabilities, which were more advanced and adaptive in AFAR-RC23. The HRI field is rapidly employing autonomous and adaptive robots that leverage AI components (e.g., ChatGPT) and capabilities to create naturalistic and smooth robot-human interactions [27]. These advancements led to improvement in the interaction outcomes and user perceptions towards the robot [27], but also to an increased complexity and reduced transparency of the robotic system, making it more difficult to control for bias. This has been supported by the literature as well [39, 65].

R5: Define field and context specific ethical principles when designing / deploying AC systems for wellbeing. Our results suggest that the use of robotic coaches may help improve fairness if, as in the case of the AFAR-RC22 [30] and AFAR-RC23 [27, 49], the design of the robot-human interactions adhere to certain ethical principles [59]. For designing robotic mental wellbeing coaches, [59] distilled a set of design and ethical recommendations through an iterative process involving stakeholders. Other relevant works [17] also highlight how the employment of ethical guideline may enhance fairness of AI-based systems [66, 67]. For example, [67] highlights the importance of ethical design in AI-based decision-making systems and [66] explores strategies to reduce bias and improve fairness in AI systems.

VIII. Discussion, Limitations and Conclusions

This work includes fairness and bias analyses of three diverse datasets. Differences in data collection protocols across these datasets have complicated systematic comparison, limiting certain analyses of bias. Future work could address this limitation by comparing datasets collected using a consistent protocol and experimental setup. For this study, we used a threshold for classifying high and low positive affect based on an older study. Future work could benefit from considering more recent data or recalculated averages of positive affect to ensure that the threshold better reflects current trends when such information becomes available. Another limitation is that we could not run statistical tests over fairness results since for each fairness score, we only have a single value over the entire test sample population. Future work can conduct more experiments with significance testing to verify the results and investigate if the proposed method is also prone to over-generalisation with high variance and its effectiveness on highly imbalanced dataset.

Moreover, the augmentation method proposed may vary in effectiveness across different datasets and HRI contexts such as resource distribution. Another limitation is that given the size of the small datasets, the variance in results would be high if different minority samples are chosen. Future work should conduct the experiments using several random augmented datasets and evaluate and report the results along with the variance in the resulting metric. Moreover, given the small sample size, it is possible that the result might be different if a different randomised augmented dataset was used. Future work can look into investigating the impact of randomised augmented dataset on the outcome of the experiments. Given our application-focused approach, we were unable to provide a thorough investigation of how our proposed method compares against other data augmentation strategies such as SMOTE which future work can look into. A comparison between the two methods will definitely provide valuable findings.

Ethical Limitations There is a huge discussion on bias and fairness in other fields such as ethics, philosophy and ML which we have been unable to thoroughly address in this paper. Moreover, different research area have been focusing on different aspects of bias and fairness which may compound the conceptual ambiguity if not appropriately defined, contextualised and discussed. As noted in Section II, majority of the fairness in wellbeing works have focused on improving fairness by reducing disparities in the ML output performance whereas most of the fairness in HRI works have primarily focused on the perception of fairness-related considerations within different HRI settings. We provide a tangential contribution of focusing on ML prediction output fairness within a HRI wellbeing coaching context which has yet to be addressed.

However, this also presents a limitation as we have primarily focused on analysing the numerical interpretation of the calculated values. These numbers should be interpreted according to the different stakeholder values and contexts [3, 6]. For instance, even though a DI ratio near 1.00 is considered more fair in our setting given the similar base rates within our experiments, an alternative fairness perspective may argue that this is less fair given that the detected prevalence of depression in females within the general population is much higher [68].

Moreover, there are several other different notions of fairness such as counterfactual fairness [69] and individual fairness [70] which we have not been able to investigate. We encourage future research to investigate these different definitions of fairness for different AC and HRI settings. We also reiterate that the interpretation of these differing notions

of fairness are context-dependent and shaped by the values of the relevant stakeholders; and neither is universally correct or incorrect. There are also other fairness considerations which we have not been able to address. For instance, it may not be considered "fair" to use perceived gender (instead of self-reported gender) as a label since gender is a self-determined identity. Future work may consider using the labels obtained from participants to avoid introducing labelling bias from external annotation [47] and continue to place emphasis on selecting the appropriate reference category in order to avoid reifying subconscious discrimination [6].

In addition, there are other fairness challenges unique to HRI that we have yet to consider. For instance, even if a ML or robot output is unbiased in the psychometric sense, it may still be perceived as unfair if one group's interactions are consistently less satisfying or they witness more interaction ruptures [48] than the other groups. We also encourage future AC and HRI works to investigate and potentially address other ethical issues or harms beyond the lenses of model bias and fairness [5].

ACKNOWLEDGMENTS

Funding: J. Cheong is funded by the Alan Turing Institute Doctoral Studentship and the Leverhulme Trust. M. Spitale is supported by PNRR-PE-AI FAIR project funded by the NextGeneration EU program. H. Gunes is supported by the EPSRC/UKRI under grant ref. EP/R030782/1 (ARoEQ). **Data access:** Raw data related to this publication cannot be openly released due to anonymity and privacy issues.

REFERENCES

[1] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," *SN computer science*, 2021.

[2] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and machine learning: Limitations and opportunities*. MIT press, 2023.

[3] S. Corbett-Davies, J. D. Gaebler, H. Nilforoshan, R. Shroff, and S. Goel, "The measure and mismeasure of fairness," *Journal of Machine Learning Research*, vol. 24, no. 312, pp. 1–117, 2023.

[4] J. Cheong, S. Kalkan, and H. Gunes, "The hitchhiker's guide to bias and fairness in facial affective signal processing: Overview and techniques," *IEEE SPM*, 2021.

[5] T. Gebru, R. Denton *et al.*, "Beyond fairness in computer vision: A holistic approach to mitigating harms and fostering community-rooted computer vision research," *Foundations and Trends® in Computer Graphics and Vision*, vol. 16, no. 3, pp. 215–321, 2024.

[6] S. S. Johfre and J. Freese, "Reconsidering the reference category," *Sociological Methodology*, vol. 51, no. 2, pp. 253–269, 2021.

[7] H. Claure, M. L. Chang, S. Kim, D. Omeiza, M. Brandao, M. K. Lee, and M. Jung, "Fairness and transparency in human-robot interaction," in *2022 HRI*. IEEE, 2022.

[8] M. L. Chang, G. Trafton, J. M. McCurry, and A. L. Thomaz, "Unfair! perceptions of fairness in human-robot teams," in *2021 RO-MAN*. IEEE, 2021.

[9] M. Brandao, M. Jirotka, H. Webb, and P. Luff, "Fair navigation planning: A resource for characterizing and designing fairness in mobile robots," *Artificial Intelligence*, 2020.

[10] A. K. Ostrowski, R. Walker, M. Das, M. Yang, C. Breazea, H. W. Park, and A. Verma, "Ethics, equity, & justice in human-robot interaction: A review and future directions," in *2022 RO-MAN*. IEEE, 2022.

[11] L. Mathur, M. Spitale, H. Xi, J. Li, and M. J. Matarić, "Modeling user empathy elicited by a robot storyteller," in *ACII*. IEEE, 2021.

[12] J. Cheong, M. Spitale, and H. Gunes, ""it's not fair!" – fairness for a small dataset of multi-modal dyadic mental well-being coaching," in *ACII 2023*, 2023.

[13] A. Bailey and M. D. Plumbley, "Gender bias in depression detection using audio features," in *EUSIPCO*. IEEE, 2021.

[14] K. Zanna, K. Sridhar, H. Yu, and A. Sano, "Bias reducing multitask learning on mental health prediction," in *ACII 2022*. IEEE, pp. 1–8.

[15] J. Cheong, S. Kuzucu, S. Kalkan, and H. Gunes, "Towards gender fairness for mental health prediction," in *IJCAI*, 2023.

[16] J. Park, R. Arunachalam, V. Silenzio, V. K. Singh *et al.*, "Fairness in mobile phone–based mental health assessment algorithms: Exploratory study," *JMIR formative research*, 2022.

[17] L. Londoño, J. V. Hurtado, N. Hertz, P. Kellmeyer, S. Voeneky, and A. Valada, "Fairness and bias in robot learning," *arXiv preprint arXiv:2207.03444*, 2022.

[18] J. Cheong, N. Churamani, L. Guerdan, T. E. Lee, Z. Han, and H. Gunes, "Causal-hri: Causal learning for human-robot interaction," in *Companion of ACM/IEEE HRI 2024*, 2024.

[19] T. Ogunyale, D. Bryant, and A. Howard, "Does removing stereotype priming remove bias? a pilot human-robot interaction study," *arXiv preprint arXiv:1807.00948*, 2018.

[20] K. S. Haring, K. Watanabe, M. Velonaki, C. C. Tossell, and V. Finomore, "Ffab—the form function attribution bias in human–robot interaction," *IEEE Transactions on Cognitive and Developmental Systems*, 2018.

[21] G. M. Alarcon, A. Capiola, I. A. Hamdan, M. A. Lee, and S. A. Jessup, "Differential biases in human-human versus human-robot interactions," *Applied Ergonomics*, 2023.

[22] C. Lachemaier, E. Lumer, H. Buschmeier, and S. Zarrieß, "Towards understanding the entanglement of human stereotypes and system biases in human-robot interaction," in *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 2024.

[23] H. Claure, Y. Chen, J. Modi, M. Jung, and S. Nikolaidis, "Multi-armed bandits with fairness constraints for distributing resources to human teammates," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020.

[24] M. L. Chang, Z. Pope, E. S. Short, and A. L. Thomaz, "Defining fairness in human-robot teams," in *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2020.

[25] M. F. Jung, D. DiFranzo, S. Shen, B. Stoll, H. Claure, and A. Lawrence, "Robot-assisted tower construction—a method to study the impact of a robot's allocation behavior on interpersonal dynamics and collaboration in groups," *ACM Transactions on Human-Robot Interaction (THRI)*, 2020.

[26] S. Jeong, L. Aymerich-Franch, K. Arias, S. Alghowinem, A. Lapedriza, R. Picard, H. W. Park, and C. Breazeal, "Deploying a robotic positive psychology coach to improve college students' psychological well-being," *UMUAI*, 2023.

[27] M. Spitale, M. Axelsson, and H. Gunes, "Vita: A multi-modal llm-based system for longitudinal, autonomous and adaptive robotic mental well-being coaching," vol. 14, no. 2, 2025.

[28] M. Axelsson, M. Spitale, and H. Gunes, "Robotic coaches delivering group mindfulness practice at a public cafe," in *Companion of the 2023 ACM/IEEE HRI*, 2023.

[29] Z. Shi, H. Chen, A.-M. Velentza, S. Liu, N. Dennler, A. O'Connell, and M. Mataric, "Evaluating and personalizing user-perceived quality of text-to-speech voices for delivering mindfulness meditation with different physical embodiments," in *HRI*, 2023, pp. 516–524.

[30] M. Spitale, M. Axelsson, and H. Gunes, "Robotic mental well-being coaches for the workplace: An in-the-wild study on form," in *HRI*, 2023.

[31] S. Jeong, L. Aymerich-Franch, S. Alghowinem, R. W. Picard, C. L. Breazeal, and H. W. Park, "A robotic companion for psychological well-being: A long-term investigation of companionship and therapeutic alliance," in *HRI*, 2023.

[32] J. Cheong, S. Kalkan, and H. Gunes, "Counterfactual fairness for facial expression recognition," in *ECCVW*. Springer, 2023.
[33] T. Schnabel and P. N. Bennett, "Debiasing item-to-item recommendations with small annotated datasets," in *ACM RecSys 2020*, 2020.
[34] J. An, L. Ying, and Y. Zhu, "Why resampling outperforms reweighting for correcting sampling bias with stochastic gradients," in *ICLR 2021*, 2021.
[35] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *ICLR 2018*, 2018.
[36] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "Augmix: A simple data processing method to improve robustness and uncertainty," in *ICLR*, 2020.
[37] C.-Y. Chuang and Y. Mroueh, "Fair mixup: Fairness via interpolation," in *ICLR 2021*, 2021.
[38] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *ICCV 2019*, 2019, pp. 6023–6032.
[39] S. Yan, D. Huang, and M. Soleymani, "Mitigating biases in multimodal personality assessment," in *ICMI*, 2020.
[40] "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, 2002.
[41] S. Wang, W. Liu, J. Wu, L. Cao, Q. Meng, and P. J. Kennedy, "Training deep neural networks on imbalanced data sets," in *2016 international joint conference on neural networks (IJCNN)*. IEEE, 2016, pp. 4368–4374.
[42] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Progress in artificial intelligence*, vol. 5, no. 4, pp. 221–232, 2016.
[43] S. Sharma, C. Bellinger, B. Krawczyk, O. Zaiane, and N. Japkowicz, "Synthetic oversampling with the majority class: A new perspective on handling extreme imbalance," in *2018 IEEE international conference on data mining (ICDM)*. IEEE, 2018, pp. 447–456.
[44] J. Chakraborty, S. Majumder, and T. Menzies, "Bias in machine learning software: Why? how? what to do?" in *ACM ESEC/FSE*, 2021.
[45] S. De Shazer and Y. Dolan, "More than miracles: The state of the art of solution-focused brief therapy," 2012.
[46] D. Watson, L. A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: the panas scales." *J. Pers. Soc. Psychol.*, 1988.
[47] J. Cheong, S. Kalkan, and H. Gunes, "Causal structure learning of bias for fair affect recognition," in *WACV 2023*, January 2023.
[48] M. Spitale, M. Axelsson, N. Kara, and H. Gunes, "Longitudinal evolution of coachees' behavioural responses to interaction ruptures in robotic positive psychology coaching," in *RO-MAN*. IEEE, 2023.
[49] M. Axelsson, M. Spitale, and H. Gunes, "" oh, sorry, i think i interrupted you": Designing repair strategies for robotic longitudinal well-being coaching," in *HRI*, 2024, pp. 13–22.
[50] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *FG*. IEEE, 2018.
[51] D. Küster, L. Steinert, M. Baker, N. Bhardwaj, and E. G. Krumhuber, "Teardrops on my face: Automatic weeping detection from nonverbal behavior," *IEEE Transactions on Affective Computing*, vol. 14, no. 4, pp. 3001–3012, 2022.
[52] E. M. Benssassi and J. Ye, "Investigating multisensory integration in emotion recognition through bio-inspired computational models," *IEEE Trans. Affect. Comput.*, 2021.
[53] M. Du, N. Liu, and X. Hu, "Techniques for interpretable machine learning," *Communications of the ACM*, 2019.
[54] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia systems*, 2010.
[55] L. Mathur and M. J. Matarić, "Introducing representations of facial affect in automated multimodal deception detection," in *ICMI*, 2020.
[56] Y. Yang, Y. Liu, and P. Naghizadeh, "Adaptive data debiasing through bounded exploration," *Adv Neural Inf Process*, 2022.
[57] W. Ma, P. Xu, and Y. Xu, "Fairness maximization among offline agents in online-matching markets," *ACM Trans. Econ. Comput.*, 2023.
[58] B. Friedman, "Value-sensitive design," *interactions*, 1996.
[59] M. Axelsson, M. Spitale, and H. Gunes, "Robots as mental well-being coaches: Design and ethical recommendations," *THRI*, 2022.
[60] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information fusion*, 2017.
[61] B. M. Booth, L. Hickman, S. K. Subburaj, L. Tay, S. E. Woo, and S. K. D'Mello, "Bias and fairness in multimodal machine learning: A case study of automated video interviews," in *ICMI*, 2021.
[62] J. Cheong, A. Bangar, S. Kalkan, and H. Gunes, "U-fair: Uncertainty-based multimodal multitask learning for fairer depression detection," *arXiv preprint arXiv:2501.09687*, 2025.
[63] Y. Wang, X. Wang, A. Beutel, F. Prost, J. Chen, and E. H. Chi, "Understanding and improving fairness-accuracy trade-offs in multi-task learning," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 1748–1757.
[64] J. Pang, J. Wang, Z. Zhu, Y. Yao, C. Qian, and Y. Liu, "Fairness without harm: An influence-guided active sampling approach," *Advances in Neural Information Processing Systems*, vol. 37, pp. 61 513–61 548, 2024.
[65] Y. Wan, W. Wang, P. He, J. Gu, H. Bai, and M. R. Lyu, "Biasasker: Measuring the bias in conversational ai system," in *Proceedings of the 31st ACM ESEC/FSE*, 2023.
[66] P. Chen, L. Wu, and L. Wang, "Ai fairness in data management and analytics: A review on challenges, methodologies and applications," *Applied sciences*, 2023.
[67] G. Biondi, S. Cagnoni, R. Capobianco, V. Franzoni, F. A. Lisi, A. Milani, and J. Vallverdú, "Ethical design of artificial intelligence-based systems for decision making," 2023.
[68] P. R. Albert, "Why is depression more prevalent in women?" pp. 219–221, 2015.
[69] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," *Advances in neural information processing systems*, vol. 30, 2017.
[70] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd innovations in theoretical computer science conference*, 2012, pp. 214–226.

Jiaee Cheong is a Turing doctoral student at the University of Cambridge. Her research interests lie at the intersection of machine learning, affective computing, fairness, causality and HRI.

Micol Spitale is an Assistant Professor at the the Politecnico di Milano, and a Visiting Affiliated Researcher at the University of Cambridge.Her research has been focused on the fields of Social Robotics, Human-Robot Interaction, and Affective Computing.

Hatice Gunes is a Full Professor of Affective Intelligence and Robotics (AFAR) in the Department of Computer Science and Technology, University of Cambridge, leading the Cambridge AFAR Lab. She is a former President of the Association for the Advancement of Affective Computing, a former Faculty Fellow of the Alan Turing Institute and is currently a Fellow of the EPSRC and Staff Fellow of Trinity Hall.