Addressing Algorithmic Bias within Facial Emotion Recognition Systems: A Human-Computer Interaction Perspective

HAILEY PARK, Statistics Department, University of Wisconsin-Madison, USA
MICHAEL F. XU, Computer Sciences Department, University of Wisconsin-Madison, USA
NAIHAO XU, Computer Sciences Department, University of Wisconsin-Madison, USA
MATTHEW BAYLES, Computer Sciences Department, University of Wisconsin-Madison, USA

Additional Key Words and Phrases: Facial emotion recognition, Algorithmic fairness/ bias, Human-Computer Interaction, Machine learning

ACM Reference Format:
Hailey Park, Michael F. Xu, Naihao Xu, and Matthew Bayles. 2024. Addressing Algorithmic Bias within Facial Emotion Recognition Systems: A Human-Computer Interaction Perspective. 1, 1 (April 2024), 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Facial Emotion Recognition (FER) systems have become deeply embedded in various aspects of our lives, as researchers and industry practitioners have found viable applications for such models in sectors ranging from healthcare [13], education [39], marketing [15], and beyond. The latest FER systems are generally able to achieve high accuracy levels when evaluated on the aggregate performances [31], and the addition of FER to existing systems and platforms have demonstrated real, practical values to their adopters.

However, as its applications become more and more wide-spread, concerns about fairness and bias issues related to these systems are also on the rise. While the aggregate performance are high, accuracy on various demographic subgroups often vary much more, where age, gender, and ethnicity have all been shown to be significant factors [11, 25, 33].

Such concerns are not unique to FER systems, as people share the same concern towards machine learning algorithms in general [19]. On the dataset side, researchers attempt to compile more balanced representations, or provide methods to synthesize new data points and augment existing datasets to become more balanced. Algorithmic attempts to address such concerns include making independent predictions for each subgroups experiencing the bias [12], making an effort to tease out and remove the bias-prone features from the learned representations [3], and so on.

While there are ongoing efforts to improve algorithmic fairness in FER systems through technical developments [6, 43], there is a lack of literature investigating human feedback to address algorithmic bias in FER systems. In this paper, we explore approaches that utilize human feedback to mitigate algorithmic bias within Facial Expression Recognition

Authors' addresses: Hailey Park, hpark353@wisc.edu, Statistics Department, University of Wisconsin-Madison, Madison, WI, USA; Michael F. Xu, michael.xu@wisc.edu, Computer Sciences Department, University of Wisconsin-Madison, Madison, WI, USA; Naihao Xu, Computer Sciences Department, University of Wisconsin-Madison, Madison, WI, USA; Matthew Bayles, Computer Sciences Department, University of Wisconsin-Madison, Madison, WI, USA.

Manuscript submitted to ACM

(FER) systems. Our focus is on examining how human input can enhance the performance and address algorithmic bias within current FER models. Our research questions are as follows:

- Is incorporating human feedback into the facial emotion recognition (FER) system effective in addressing algorithmic bias?
- Does the intervention of human feedback result in a trade-off of model accuracy?
- How can we incorporate human feedback to improve FER systems in terms of both algorithmic fairness and model performance?

2 RELATED WORKS

2.1 Fairness and Bias

As Artificial Intelligence (AI) and Machine Learning (ML) algorithms are increasingly integrated into various facets of our daily lives, concerns about algorithmic fairness and bias have surfaced concurrently. However, there aren't any standard definitions or measures universally used to determine whether an algorithm is biased. There are several different notions to measure whether an algorithm is "biased" with varying advantages and disadvantages[17, 34]. Hence, it is crucial to consider the proper legal, ethical, and social contexts to determine what measure is appropriate to measure the algorithmic bias [34]. Here, we introduce some common measures of algorithmic fairness frequently used in the context of ML fairness literature on classification tasks.

Disparate impact [14] is the framework used in a legal setting to quantitatively measure an unintentional bias that can occur in a selection process in businesses. Formally, disparate impact can be computed as follows:

$$\frac{P(\hat{Y} = 1|A \neq 1)}{P(\hat{Y} = 1|A = 1)} \geq 1 - \epsilon$$

$\hat{Y} = 1$ denotes that an algorithm has made a correct prediction in a classification task, and $A \neq 1$ represents an underrepresented group attribute and $A = 1$ representing a majority group attribute. An attribute herein refers to a protected class like race or gender. The goal of removing disparate impact is to change the remaining non-sensitive attributes for the algorithm's classification performance to not be disproportionate across different groups, while keeping the overall performance as high as possible [14]. Intuitively, an algorithm with disparate impact will have a much high ratio achieving correct prediction in a majority class of a certain protected attribute A in comparison to a minority class. However, one downside of demographic parity is that a classifier may be considered unfair if the base rates corresponding to two different classes are inherently different [34].

Equalized odds addresses this disadvantage by computing the difference in false-positive rates (FPR) and true-positive rates (TPR) [18].

$$|P(\hat{Y} = 1|A = 1, Y = y) - P(\hat{Y} = 1|S \neq 1, Y = y)| \leq \epsilon, y \in \{0, 1\}$$

Under this definition of equalized odds, an algorithm is considered fair if the difference between FPR and TPR between the majority and minority groups is small.

Equal calibration [8] is another notion of fairness that has been shown to be incompatible with the equalized odds [35]. Assessing calibration is especially considered important inthe context of designing or auditing a risk assessment

algorithm [10]. Equal calibration is formulated as follows:

$$|P(Y = 1|A = 1, V = v) - P(Y = 1|A \neq 1, V = v)| \leq \epsilon$$

V denotes the predicted probability value. Under this measure, an algorithm is considered fair if for any predicted probability value $V = v$, both the majority group and the minority group have roughly the same probability of belonging to the true positive class $Y = 1$)[42]. A major distinction between the previous two notions is that this definition considers the true probability.

It has been shown, however, that this notion of calibration and error-rate fairness measures are inherently at odds [35]. In light of these findings, an appropriate measure to quantify algorithmic bias should be chosen by accounting for the model's legal, social, and ethical contexts [34]. Formal mathematical measures of algorithmic fairness can be helpful in quantifying the potential bias within the model and to adjust the model accordingly. However, there are some findings that these popular measures of algorithmic fairness can harm the underrepresented groups that these measures were designed to protect [10]. Furthermore, a prior study documents that there is an inherent trade-off between the model's prediction accuracy and algorithmic fairness [27]. That is, it is challenging to maintain the model's prediction performance as we make adjustments to improve algorithmic fairness. To address these concerns, there has been a move to adopt a consequentialist framework in measuring algorithmic bias[7]. This approach considers the consequences of decisions from the stakeholder's perspective instead of measuring algorithmic fairness based on formal definitions.

2.2 Quantifying FER Performances

FER metrics have largely stayed the same over the last five years. Most models are trained on datasets that express a set of emotions. The most common emotions are anger, sadness, happiness, fear, disgust, and surprise[30]. These emotions are widely believed to be universally recognized by all humans regardless of culture [37]. As a result, these emotions are widely used as the benchmark to determine how accurate a model can determine emotion. Accuracy, precision, and specificity are some of the metrics that are used as benchmarks[9].

2.3 Improving FER model Performances

With the rapid increase in power and accessibility to machine learning FER systems are becoming increasingly common[23]. Specifically, models designed using Convolutional Neural Networks (CNN) are at the heart of this movement but there are still algorithms that use more traditional approaches such as support vector machines. However, these traditional approaches lack the generality that is potentially possible with a CNN[5]. Four years ago some of the first models to use facial emotion recognition using convolutional neural network (FERC) had a success rate of around 96% [30] were created. They used their model to removed the background on images to help reduce noise. Despite this high early success rate some datasets proved easier to model than others. Depending on the model researchers were getting between 65% to 99 % accuracy in 2020 [32].

However other the next couple of years model progress continued to progress. For example a popular dataset,JAFFE, saw model accuracy clime from 96% to 99% in less than one year[2]. Another dataset, FER2013, saw model accuracy increase from 65% to 75% in two years' time [23] as researchers created new models and modified existing models.

One of current issues with using CNNs is the datasets themselves. Most of the datasets researchers have used to train their models are essential laboratory level images that express one specific emotion. In the real world, however, humans often express many different emotions and combinations of emotions but models developed within the last year are still using datasets that only define six emotions [21]. Additionally, there are other real world issues such as image

corruption due to compression and how the face is framed in the image. A recent study incorporated purposeful errors into their image datasets and tested four of the most popular models and found that the error rate of accepted models can increase by 70% to 200% depending on the model[16]. This shows that there is still plenty of room to improve FER performance.

2.4 Improving FER model Fairness

With the advent of commercialization of machine learning algorithms, more and more attention has been given to their fairness performance in practice. Through anonymous surveys and semi-structured interviews, Holstein et al. [19] highlighted the need for more proactive and holistic auditing methods for fairness in machine learning systems. Specifically, they've argued that even with humans in the loop, more intuitive tools are needed to mitigate the biases that may be embedded in humans. In a regression setting, Yan et al. [46] proposed a multi-layer factor analysis framework to identify heterogeneity patterns within the dataset, and demonstrated that when paired with feature rescaling based on the factor analysis results, they were able to improve model fairness without directly accessing sensitive attributes of the data such as race and gender. In the context of FER, Xu et al. [45] investigated bias and fairness issues by applying variations of both a "fairness through awareness" approach [12], and a "fairness through unawareness" approach [3, 44], which they've referred to as their attribute-aware approach, and disentangled approach, respectively. They found that data augmentation to improve sample balance does help mitigating bias, but that both attribute-aware and disentangled approaches were able to further improve the fairness performance of the baseline algorithm.

Kara et al. [22] proposed a Continual Learning framework as a mitigation strategy to enhance fairness in FER systems. They've demonstrated promising results in terms of achieving higher Fairness Scores while maintaining high accuracy, compared to non-Continual Learning methods.

2.5 Existing Datasets and Models

As facial recognition becomes more and more popular in a variety of areas, how to come up with models better overall performances and keys parts to form effective datasets have raised lots of attention. Most of the existing approaches perform tests on seven major categories of emotions: **anger**, **disgust**, **fear**, **happiness**, **sadness**, **surprise**, and **neutral** and attempts to increase the prediction accuracy compared to the existing state-of-the-art methods.

Currently, here are lots of datasets available for use; for example, FER-2013 by Kaggle and FER+ by Microsoft. But doubts exist as well on the effectiveness of such datasets. Kim et al. [26] utilized the FACES dataset on commercial FER systems: **Amazon Rekognition**, **Face++**, **Microsoft**, and **Sighthound**, which are the most popular commercial FER systems. They specifically focused on six emotions: **anger**, **disgust**, **fear**, **happiness**, **neutrality**, and **sadness** and compared accuracy across ages and genders. The result indicates a potential age bias for FACES dataset: as ages increase, the average accuracy decrease for both genders, and the overall accuracy for both genders is tight. Kim et al. [26] conclude that datasets like FACES are lack of large samples from different racial groups. They have also pointed that lots of existing datasets are lack of diversity and natural expressions. For example, datasets like AffectNet or EmotioNet, often have better demographic representation but have limited information on unidentified expressions. Increasing the diversity and classifying more expressions are the keys to build better datasets instead of simply adding more training data for under-represented subgroups.

Face Valuing is an approach introduced by Veeriah et al. [41]. This grip-selection based approach is designed to train learning agent to interpret user's facial expressions in front of a webcam by adjusting to user's preferences; thus, reducing less direct feedback. There are 68 key points, which are 2-D coordinates, from a frame containing a human's

face being detected through a popular facial landmark detection algorithm as cited in Veeriah et al. [41]. These key points are normalized from each frame and meant to denote the position of certain special locations of a human's face, and 23 of them are selected correspond to the positions of eye brows and mouth of a human's face, which produced sufficient variations between different facial expressions. This approach reduces the total number of expensive human generated-rewards, or simply saying the total running time substantially varied from 30% to 50% based upon the given number of objects and grips compared to the trials without face valuing.

The **Reinforcement learning for pre-selecting useful images** (RLPS) is a novel framework based on reinforcement learning for pre-selecting useful images proposed by Li and Xu [28]. It is consisted of two steps: a image selector and a rough emotion classifier. The former is used to select useful images for emotion classification through reinforcement strategy whereas the latter acts like a teacher to help train image selector. Unlike other existing approaches, which focus on getting high-quality datasets, RLPS works like an add-on algorithm, which corporates with the existing classifiers. It is designed to improve classification performance by increasing the quality of the dataset and make it available to be applied to any classifier. **RAF-DB**, **ExpW**, and **FER2013** are three datasets that tests performed on in this article. Noise data were also added when performing tests. When combined with DCNN method, RLPS performs better than the DCNN baseline. Results vary from 2% to 10% when performing on different datasets. Moreover, the RLPS approach makes extraordinary improvements on fear and disgust emotions when tests are performed in RAF-DB and ExpW datasets respectively. Future work will focus on noisy environments and make it less dependent on the initial classifiers.

The **Discriminative Deep multi-task learning** (DDMTL), is an enhanced version of the previous DMTL algorithm proposed by Zheng et al. [48]. It aims to overcome the problem that DMTL only considers the information of class labels, while ignoring the local information of sample spatial distribution. Tests have been performed on three datasets: **extended Cohn-Kanade dataset** (CK+), the **MMI facial expression dataset** (MMI), and the **Static Facial Expressions in the Wild dataset** (SFEW), and results are compared to different existing state-of-the-art methods, such as **AlexNet**, **VGGNet**, **GoogleNet**, **AdaGabor**, **3D-CNN**, **SJMT**, and **DMTL** on all three datasets. Test results from all three datasets indicate that the new DDMTL approach has successfully improved the accuracy as expected. It has approximately 3% increased accuracy compared to the old version of DMTL and high up to 8% when comparing to other state-of-the-art methods. Future direction suggested from the article is to focus on maintaining the recognition accuracy when dealing with combined facial expression datasets.

Another outstanding algorithm which increases the overall accuracy is called **Region-based Convolutional Fusion Network** (RCFN) by Ye et al. [47] This approach builds a muscle movement model which takes user's frontal face and extract forehead, eye, and mouth patches as crucial regions to remove the unrepresentative regions and interferences by facial organs. Then, a fast and practical network is designed to extract robust triple-level features from low to semantic level in each crucial region and fuse them for FER. Lastly, a constrained punitive loss is used for network optimization by automatically adjusting the loss function according to the output to improve the FER performance. From the given test results, RCFN performs well in commonly used datasets such as **KDEF**, **CK+**, and **Oulu-CASIA**. It has an average of 5% increased prediction accuracy and takes an average of 50% less analysis time compared to other popular approaches; for instance, WCFN and CCFN.

3 METHODS

We had initially surveyed algorithmic bias and solutions to reduce its impact in FER systems in a prior milestone. To conduct a study with further emphasis on the "human" component of the HCI, we had reoriented our research direction to investigate personalized FER systems.

We aim to understand how we can design personalized FER system through user feedback. Specifically, we study whether gathering more user-specific data through the deployment of the system can improve the accuracy of the FER model. Furthermore, we study how we can collect this data and user feedback without disrupting the original purpose of the system. We will employ two methodologies to investigate each aspects of personalized FER system design. First, we conduct an experiment that empirically tests whether gathering user-specific data can improve the model performance. Then, through a web-based questionnaire, we investigate user preferences in terms of feedback collection mechanism employed by such systems.

3.1 "Personalized" FER Model through User Feedback

Is collecting user-specific data effective in improving the FER model through adding personalizing touch to a pre-trained model? We investigate this by utilizing a FER system based on Deep Convolutional Neural Network (DCNN) architecture and Transfer Learning (TL). DCNN models achieve the state-of-the art accuracy in facial emotion reognition tasks [24, 36]. Additionally, transfer learning framework enables us to augment the pre-trained model with additional pictures of a user. Some studies have shown that transfer learning can also enhance the performance of the pre-trained network [38].

3.1.1 System Design. We utilize Transfer Learning technique to integrate user-specific images into the pre-trained dataset. We'll emulate the model architecture from a prior study that employed transfer learning in FER model [1]. We deem this system appropriate because the model's facial emotion recognition task aligns with our primary focus and the observed high model performance. We'll expand this system by crafting a dataset comprising facial images of an individual. This aims to replicate the hypothetical FER system product that collects user images while the system is in active use. JAFFE (Japanese Female Facial Expression) is a widely employed dataset in FER model training, comprising 213 images from 10 different Japanese female subjects. We'll partition this dataset into subsets containing unique individuals as additional data to fine-tune the pre-trained network. We select this dataset because it's the most widely used FER dataset with the fewest unique individuals, making it easier to subgroup the original data.

3.1.2 System Evaluation. We will compare model performance before any Transfer Learning is applied to the model with model performance after transfer learning is applied. Each participant will sit in front of a computer equipped with a webcam and a program with a basic user interface (UI) designed by the research team. The researcher conducting the session will start the program and sit off to the side. The session can be done in person or virtually. The UI will display pictures and prompt the user to select which emotion they are experiencing while viewing the image. If the session is virtual the researcher will select the emotion for the participant. The emotions are anger, fear, disgust, happiness, and sadness. A group of two sets of sixteen images will be selected by the research team. One set for each stage of the experiment. Each emotion being studied will have four images associated with it. The images will be shown in a random order with every image being shown exactly one time. The response of the participant and the response of the model will be recorded by the program. While the participant is being shown the images, the model will capture an image of the participant experiencing emotion using the webcam for each image they see, if the session is in person. If the session is virtual the researcher will record the session and the participant images will be extracted at later data. These images will be used to predict the participants emotion and will be used in the transfer model. Using the user's input, the taken image and their response will be applied to the model using Transfer Learning. Once the whole series of images and responses have been captured, we will run the same procedure as above but with the updated model and

a different set of sixteen images. Depending on the training time of the model and whether the session is virtual this can either occur in the same session or later.

3.1.3 System Analysis. To measure performance, we will apply a statistical analysis to the participants's feedback compared with the predicted response of the model and analyze model statistics the research team deems important such as model training time. Our pool of participants will consist of twenty adult individuals who are friends or family of the research team.

3.2 User Preferences on Feedback Mechanism

In this subsection of our work, we aim to explore the following questions:

(1) **RQ 1:** What is the preferred channel and mechanism for users to provide feedback to an automated FER system?
(2) **RQ 2:** Does the preference depend on the deployment context of the FER system?

Studies in human-AI interaction has demonstrated that the field is complex and multi-dimensional. Among other things, the role the system assumes [29], the proactiveness of the system [40], and the modality of the feedback [20], are all important considerations. For this project, our goal is to understand the distribution of user preferences over these various dimensions, specifically in the context of human feedback for FER systems.

3.2.1 Study Setting and Population. As there is limited existing research on this specific topic, we conducted an exploratory questionnaire with a total of 38 adult participants, consisting of friends and families of the research team. This population is generally tech-savvy and familiar with the various applications of systems involving digital faces, such as Apple's Face ID and conference calls, but may not be entirely familiar with the FER system itself (average self-reported familiarity with FER systems is around 25 on a scale of 0 to 100). More characteristics of the participants are shown in Table 1.

While our research question is generally applicable to scenarios where FER systems are intimately positioned along with human users, for this specific exploratory questionnaire, to help participants contextualize our questions and their responses, we will describe two specific settings involving FER systems: 1. An in-home social robot equipped with a camera for vision tasks, and 2. A screen-based application such as conference call software. Participants will be randomly asked about their preferences in one of these settings. For ease of comparison, the sets of questions for each scenario is essentially equivalent, except for some terms referring to the robot vs. the application.

3.2.2 Study Design. In order to answer the research questions, especially *RQ 2*, we employed a between-subject design, where participants are randomly assigned one of the two deployment scenarios (robot *vs.* screen-based apps). To better control for individual-level variances, it would have been preferable to conduct a within-subject study, with the order of the scenarios counter balanced over the participants. However, our internal testing as well as pilot testers revealed that having participants complete both scenarios would have been too much of a burden, which could negatively impact the quality of data we receive. For these reasons, we've opted for the between-subject design. Along with this choice, we also decided that instead of the original target of 20 participants, we will aim to recruit 40 participants to increase the power of the analysis.

3.2.3 Data Collection. The questionnaire was created based on existing literature in related fields, and refined through workshop sessions among the research team. We asked participants to evaluate both high-level preferences such as modality and frequency of feedback solicitation, and low-level preferences such as comparisons of several variations of

Table 1. Demographic Information of Participants

	Frequency	Percentage
Gender		
Male	19	50.0
Female	19	50.0
Age Range		
16 - 20	2	5.26
21 - 25	12	31.58
26 - 30	3	7.89
Over 30	21	55.26
Education		
High School	2	5.26
Some College	6	15.79
Bachelor's	15	39.47
Master's	13	34.21
Doctoral	2	5.26

a feedback mechanism. We also included open-ended questions to allow participants to mention other aspects they deem important, but not covered in the questionnaire, and to elaborate on their responses. After several brainstorming sessions within the research team, we settled on a finalized version of the questionnaire, digitalized and deployed it on a web-based survey platform, Qualtrics. Links to the anonymous questionnaire were distributed to participants, who then completed the questionnaire on Qualtrics, where the data are collected and stored.

3.2.4 Questionnaire Structure and Sample Questions. The questionnaire starts with a set of standard demographic questions, where we also asked participants about their familiarity with FER systems. The participants are then randomly assigned one of the two scenarios, an FER system embedded on a domestic robot, and an FER system running on a standard screen-based application (e.g. conference app). In each case, the participants are first introduced to the scenario, where potential utility and use cases of FER are described to the users. Next, participants are asked to rate a series of questions on sliders, with a scale from 0 to 100. An example question is "*In terms of providing feedback to the software's emotion readings, do you prefer that the software asks for feedback in a synchronous or asynchronous fashion?*" where we also provided a description of what each option would be reflected in the scenario. Another example is "If the software needs to ask for feedback, how do you prefer that it's initiated?" For this question, we included three options, each paired with its own slider for the participants to express their preferences. The three options are *Audio: Verbally asks through natural language*, *Visual: Text on display screen*, and *Visual & Audio: Text on display screen & a chime / "ding" sound.* Additionally, we've also touched on aspects such as the frequency of interaction, and preferences regarding ways to improve or customize such interactions. Throughout the questionnaire, we also included opportunities for open responses, and encourage participants to provide more context or suggest other interaction modalities or considerations.

3.2.5 Data Analysis. We will perform both qualitative and quantitative analysis on the survey responses. We will generate descriptive statistics and analyze for patterns for the quantitative sections. For the qualitative section, we plan to conduct a Thematic Analysis [4]. We will first generate potential codes, and then individually assign codes to the relevant responses. We will then discuss and iterate on the code assignment until an agreement is reached for each response.

REFERENCES

[1] MAH Akhand, Shuvendu Roy, Nazmul Siddique, Md Abdus Samad Kamal, and Tetsuya Shimamura. 2021. Facial emotion recognition using transfer learning in the deep CNN. *Electronics* 10, 9 (2021), 1036.

[2] M. A. H. Akhand, Shuvendu Roy, Nazmul Siddique, Md Abdus Samad Kamal, and Tetsuya Shimamura. 2021. Facial Emotion Recognition Using Transfer Learning in the Deep CNN. *Electronics* 10, 9 (2021). https://doi.org/10.3390/electronics10091036

[3] Mohsan Alvi, Andrew Zisserman, and Christoffer Nellåker. 2018. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. 0–0.

[4] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[5] Felipe Zago Canal, Tobias Rossi Müller, Jhennifer Cristine Matias, Gustavo Gino Scotton, Antonio Reis de Sa Junior, Eliane Pozzebon, and Antonio Carlos Sobieranski. 2022. A survey on facial emotion recognition techniques: A state-of-the-art literature review. *Information Sciences* 582 (2022), 593–617. https://doi.org/10.1016/j.ins.2021.10.005

[6] Yunliang Chen and Jungseock Joo. 2021. Understanding and mitigating annotation bias in facial expression recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14980–14991.

[7] Alex Chohlas-Wood, Madison Coots, Henry Zhu, Emma Brunskill, and Sharad Goel. 2021. Learning to be fair: A consequentialist approach to equitable decision-making. *arXiv preprint arXiv:2109.08792* (2021).

[8] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.

[9] M Kalpana Chowdary, Tu N Nguyen, and D Jude Hemanth. 2023. Deep learning-based facial emotion recognition for human–computer interaction applications. *Neural Computing and Applications* 35, 32 (Nov. 2023), 23311–23328.

[10] Sam Corbett-Davies, Johann Gaebler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel. 2023. The measure and mismeasure of fairness. *arXiv preprint arXiv:1808.00023* (2023).

[11] Iris Dominguez-Catena, Daniel Paternain, and Mikel Galar. 2022. Gender Stereotyping Impact in Facial Expression Recognition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 9–22.

[12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[13] Zixiang Fei, Erfu Yang, David Day-Uei Li, Stephen Butler, Winifred Ijomah, Xia Li, and Huiyu Zhou. 2020. Deep convolution network based emotion analysis towards mental health care. *Neurocomputing* 388 (2020), 212–227.

[14] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 259–268.

[15] M Rosario González-Rodríguez, M Carmen Díaz-Fernández, and Carmen Pacheco Gómez. 2020. Facial-expression recognition: An emergent approach to the measurement of tourist satisfaction through emotions. *Telematics and Informatics* 51 (2020), 101404.

[16] Antonio Greco, Nicola Strisciuglio, Mario Vento, and Vincenzo Vigilante. 2023. Benchmarking deep networks for facial emotion recognition in the wild. *Multimedia Tools and Applications* 82, 8 (March 2023), 11189–11220.

[17] Swati Gupta, Akhil Jalan, Gireeja Ranade, Helen Yang, and Simon Zhuang. 2020. Too many fairness metrics: Is there a solution? *Available at SSRN 3554829* (2020).

[18] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016).

[19] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–16.

[20] Ananya Ipsita, Hao Li, Runlin Duan, Yuanzhi Cao, Subramanian Chidambaram, Min Liu, and Karthik Ramani. 2021. VRFromX: from scanned reality to interactive virtual experience with human-in-the-loop. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–7.

[21] Deepak Kumar Jain, Ashit Kumar Dutta, Elena Verdú, Shtwai Alsubai, and Abdul Rahaman Wahab Sait. 2023. An automated hyperparameter tuned deep learning model enabled facial emotion recognition for autonomous vehicle drivers. *Image and Vision Computing* 133 (2023), 104659. https://doi.org/10.1016/j.imavis.2023.104659

[22] Ozgur Kara, Nikhil Churamani, and Hatice Gunes. 2021. Towards fair affective robotics: continual learning for mitigating bias in facial expression and action unit recognition. *arXiv preprint arXiv:2103.09233* (2021).

[23] Yousif Khaireddin and Zhuofa Chen. 2021. Facial Emotion Recognition: State of the Art Performance on FER2013. *CoRR* abs/2105.03588 (2021). arXiv:2105.03588 https://arxiv.org/abs/2105.03588

[24] Yousif Khaireddin and Zhuofa Chen. 2021. Facial emotion recognition: State of the art performance on FER2013. *arXiv preprint arXiv:2105.03588* (2021).

[25] Eugenia Kim, De'Aira Bryant, Deepak Srikanth, and Ayanna Howard. 2021. Age bias in emotion detection: An analysis of facial emotion recognition performance on young, middle-aged, and older adults. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 638–644.

[26] Eugenia Kim, De'Aira Bryant, Deepak Srikanth, and Ayanna Howard. 2021. Age Bias in Emotion Detection: An Analysis of Facial Emotion Recognition Performance on Young, Middle-Aged, and Older Adults. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*

(Virtual Event, USA) *(AIES '21)*. Association for Computing Machinery, New York, NY, USA, 638–644. https://doi.org/10.1145/3461702.3462609

[27] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).

[28] Huadong Li and Hua Xu. 2020. Deep reinforcement learning for robust emotional classification in facial expression recognition. *Knowledge-Based Systems* 204 (2020), 106172. https://doi.org/10.1016/j.knosys.2020.106172

[29] Mengqi Liao and S Shyam Sundar. 2021. How should AI systems talk to users when collecting their personal information? Effects of role framing and self-referencing on human-AI interaction. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[30] Ninad Mehendale. 2020. Facial emotion recognition using convolutional neural networks (FERC). *SN Applied Sciences* 2, 3 (Feb. 2020), 446.

[31] Wafa Mellouk and Wahida Handouzi. 2020. Facial emotion recognition using deep learning: review and insights. *Procedia Computer Science* 175 (2020), 689–694.

[32] Wafa Mellouk and Wahida Handouzi. 2020. Facial emotion recognition using deep learning: review and insights. *Procedia Computer Science* 175 (2020), 689–694. https://doi.org/10.1016/j.procs.2020.07.101 The 17th International Conference on Mobile Systems and Pervasive Computing (MobiSPC),The 15th International Conference on Future Networks and Communications (FNC),The 10th International Conference on Sustainable Energy Information Technology.

[33] Jaspar Pahl, Ines Rieger, Anna Möller, Thomas Wittenberg, and Ute Schmid. 2022. Female, white, 27? bias evaluation on data and algorithms for affect recognition in faces. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 973–987.

[34] Dana Pessach and Erez Shmueli. 2022. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)* 55, 3 (2022), 1–44.

[35] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. *Advances in neural information processing systems* 30 (2017).

[36] E Pranav, Suraj Kamal, C Satheesh Chandran, and MH Supriya. 2020. Facial emotion recognition using deep convolutional neural network. In *2020 6th International conference on advanced computing and communication Systems (ICACCS)*. IEEE, 317–320.

[37] J A Russell. 1994. Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. *Psychol Bull* 115, 1 (Jan. 1994), 102–141.

[38] Manali Shaha and Meenakshi Pawar. 2018. Transfer learning for image classification. In *2018 second international conference on electronics, communication and aerospace technology (ICECA)*. IEEE, 656–660.

[39] Güray Tonguç and Betul Ozaydın Ozkara. 2020. Automatic recognition of student emotions from facial expressions during a lecture. *Computers & Education* 148 (2020), 103797.

[40] Niels Van Berkel, Mikael B Skov, and Jesper Kjeldskov. 2021. Human-AI interaction: intermittent, continuous, and proactive. *Interactions* 28, 6 (2021), 67–71.

[41] Vivek Veeriah, Patrick M. Pilarski, and Richard S. Sutton. 2016. Face valuing: Training user interfaces with facial expressions and reinforcement learning. *CoRR* abs/1606.02807 (2016). arXiv:1606.02807 http://arxiv.org/abs/1606.02807

[42] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the international workshop on software fairness*. 1–7.

[43] Yiming Wang, Hui Yu, Brett Stevens, and Honghai Liu. 2015. Dynamic facial expression recognition using local patch and lbp-top. In *2015 8th International conference on human system interaction (HSI)*. IEEE, 362–367.

[44] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. 2020. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8919–8928.

[45] Tian Xu, Jennifer White, Sinan Kalkan, and Hatice Gunes. 2020. Investigating bias and fairness in facial expression recognition. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*. Springer, 506–523.

[46] Shen Yan, Hsien-Te Kao, Kristina Lerman, Shrikanth Narayanan, and Emilio Ferrara. 2021. Mitigating the bias of heterogeneous human behavior in affective computing. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 1–8.

[47] Yingsheng Ye, Xingming Zhang, Yubei Lin, and Haoxiang Wang. 2019. Facial expression recognition via region-based convolutional fusion network. *Journal of Visual Communication and Image Representation* 62 (2019), 1–11.

[48] Hao Zheng, Ruili Wang, Wanting Ji, Ming Zong, Wai Keung Wong, Zhihui Lai, and Hexin Lv. 2020. Discriminative deep multi-task learning for facial expression recognition. *Information Sciences* 533 (2020), 60–71. https://doi.org/10.1016/j.ins.2020.04.041