Gender Stereotyping Impact in Facial Expression Recognition

Iris Dominguez-Catena$^{[0000-0002-6099-8701]}$, Daniel Paternain$^{[0000-0002-5845-887X]}$, and Mikel Galar$^{[0000-0003-2865-6549]}$

Institute of Smart Cities (ISC), Department of Statistics, Computer Science and Mathematics
Public University of Navarre (UPNA)
Arrosadia Campus, 31006, Pamplona, Spain
{iris.dominguez, mikel.galar, daniel.paternain}@unavarra.es

Abstract. Facial Expression Recognition (FER) uses images of faces to identify the emotional state of users, allowing for a closer interaction between humans and autonomous systems. Unfortunately, as the images naturally integrate some demographic information, such as apparent age, gender, and race of the subject, these systems are prone to demographic bias issues. In recent years, machine learning-based models have become the most popular approach to FER. These models require training on large datasets of facial expression images, and their generalization capabilities are strongly related to the characteristics of the dataset. In publicly available FER datasets, apparent gender representation is usually mostly balanced, but their representation in the individual label is not, embedding social stereotypes into the datasets and generating a potential for harm. Although this type of bias has been overlooked so far, it is important to understand the impact it may have in the context of FER. To do so, we use a popular FER dataset, FER+, to generate derivative datasets with different amounts of stereotypical bias by altering the gender proportions of certain labels. We then proceed to measure the discrepancy between the performance of the models trained on these datasets for the apparent gender groups. We observe a discrepancy in the recognition of certain emotions between genders of up to 29% under the worst bias conditions. Our results also suggest a safety range for stereotypical bias in a dataset that does not appear to produce stereotypical bias in the resulting model. Our findings support the need for a thorough bias analysis of public datasets in problems like FER, where a global balance of demographic representation can still hide other types of bias that harm certain demographic groups.

1 Introduction

The development of technology in the last decades, especially in Machine Learning (ML) and Artificial Intelligence (AI), has exposed an ever-growing portion of the population to autonomous systems. These systems, from the mundane autocorrector in mobile devices to the critical autopilot in self-driving cars, impact the lives of people around the world. Despite their continuous improvement in all respects, this impact is not always positive. A point of particular concern is when the mistakes our AI systems make systematically harm certain demographic groups. We call this behavior an unwanted bias. Unwanted biases can be based on several demographic characteristics, the most common being age, sex, and race [19,25,10].

These biases have been studied and classified into many types according to the stage of the ML life cycle from which they originate [29]. Although all sources of bias must be taken into account to develop fair systems, dataset bias has gained special relevance in the last decade. For many ML applications, Deep Learning algorithms that use large amounts of data have become the standard approach [6]. This has led to the creation of large public datasets and to the decoupling of the dataset creation and model training phases. These datasets, despite their usefulness, many times exhibit heavy biases [25] that are easy to overlook for the teams using them. These dataset biases can be found in the demographic proportions of the datasets [18], in the relationships between data of multimodal datasets [33,8], in the sample labeling and the label themselves [25], and even in the images of the dataset [31]. A specific type of bias, the topic of this work, is stereotypical bias [1], where demographic groups can be equally represented but over or underrepresented in certain categories.

The impact of these biases on the predictions of the final model is highly variable, depending on both the severity and nature of the biases and the context of the application itself. For applications that involve human users, especially when the implementation of the system regulates access to resources or involves the representation of people, unfair predictions can directly lead to harm to population groups [19,5] (allocative and representational harms).

A current area of interest in AI is Facial Expression Recognition (FER) [22]. FER refers to a modality of automatic emotion recognition in which, from a picture of a face, the system predicts the emotional state of the subject. The readiness for implementation, possible with minimal hardware, combined with the nature of the data involved, makes FER an application where biases are easily developed and could potentially lead to representational harms. Furthermore, the face images have some demographic information naturally integrated into them, such as apparent age, gender, or race. With most datasets lacking explicit external demographic labels, bias mitigation techniques are hard to apply, and even bias detection poses a challenge. Regarding gender in particular, although public FER datasets are usually globally balanced, with similar proportions of male and female presenting people, they often hide stereotypical biases. That is, they are unbalanced for certain categories, despite the global balance, which can systematically skew the final model predictions depending on the subject's apparent gender.

In this work, we analyze stereotypical gender bias in the context of FER. In particular, we focus on the FER+ dataset [6], a refined version of the popular FER2013 dataset [16]. With this dataset as the base, in our experiments we generate derivative datasets with different amounts of stereotypical bias by altering the gender proportions of certain labels and measuring the variations in the final model predictions. These induced biases allow us to quantify the limits of the variations under extreme stereotypical bias conditions in FER problems. Although previous work [2,32,14,12,13] has also studied and revealed biases in general and gender biases in particular in FER, no other work has focused on the problem of stereotypical biases in this context. We hope that our contribution will help establish the importance of this type of bias and understand the extent of its impact in this context.

From our results, we observe a discrepancy in the recognition of certain emotions between genders of up to 29% under the worst bias conditions. Our results also suggest a safety range for stereotypical bias in the dataset that does not appear to produce bias in the final model. These findings can help future implementations avoid some potential harms in FER due to misrepresentation of groups.

The remainder of this work is organized as follows. Section 2 describes the related work and some background information for our proposal. Next, Section 3 describes the proposed experiments and the relevant implementation details. In Section 4 presents and analyzes the results of the experiments. Finally, Section 5 concludes this work and proposes future work.

2 Related work

2.1 Facial Expression Recognition

FER is one of the simplest and most widespread modalities of the more general automatic emotion recognition. In automatic emotion recognition, the system tries to identify the emotional state of a person from their expressions and physiology. Several modalities are possible, depending on both the input data required by the system and the output codification of the emotional state [3]. FER, in particular, uses as input data a static image or a video of a human face, making it relatively easy to deploy with minimal hardware.

Regarding the emotion codification, the classical approaches are continuous [23] and discrete models of emotion [15]. The continuous model separates emotion into several independent dimensions, such as *valence* and *arousal*. Instead, the discrete model assimilates emotions into several prototypes, with the most common categorization being the six basic emotions of Ekman [15]: angry, disgust, fear, happiness, sadness, and surprise. Although the continuous codification is more expressive, the labeling of samples is more subjective and complex. Thus, most FER datasets are based on the discrete approach. In this work, we will focus on the same discrete approach.

2.2 Bias

Most definitions of fairness are based on the idea of absence of unwanted bias [30]. This unwanted bias, understood as a systematic variation in the treatment of a demographic group that can potentially lead to harm.

Although most definitions of bias as a proxy for fairness are designed around the predictions of a model, the general concept of bias can be linked to different sources of bias at different points of the ML life cycle [29]. In particular, for applications in ML where public datasets are common, bias present in the source data is particularly relevant [24]. Large datasets, in particular, have been subject to extensive analysis, finding different types of bias [25,11].

While data bias is predominantly studied in the form of representational bias, where certain demographic groups are overly prevalent in a dataset, another common bias in some types of datasets is *stereotypical bias* [1,9]. In classification tasks, this kind of bias is modeled as a correlation between the demographic attributes of a subject and the problem classes, and can easily leak into the datasets as different demographic profiles for certain classes.

Some works have already analyzed FER systems, finding demographic bias in general [20,17], including several instances of gender biases [2,14,12,13]. In particular, Ahmad et al. [2] analyzes the prediction of commercial systems, without working with the bias in the original datasets. Domnich and Anbarjafari [14] study the gender bias exhibited by six different neural networks trained for FER. Deuschel et al. [12] employ intentionally biased datasets, composed only of male or female subjects, to study the impact of these biases on the detection of action units, a problem closely related to FER. Finally, Dominguez et al. [13] also uses intentionally biased and balanced datasets to validate a set of metrics for bias detection, using FER as a case study and showing inherent representational and stereotypical biases in some FER datasets.

Unlike the previous work, we will focus on the stereotypical bias in FER. We will employ progressively biased datasets to measure the impact of this type of bias on the trained model. Bias is often measured with specific bias metrics, which helps quantify its impact. Despite this, it is important to notice that any application of a specific metric still requires a proper qualitative discussion of its context, or it can easily lose its usefulness [26]. For this reason, in this work, we will employ a qualitative and intuitive approach without employing a specific metric. Nonetheless, we will look for deviations in recalls (accuracy constrained to the examples of a certain class) between demographic groups, with an underlying notion of fairness consistent with the *conditional use accuracy equality* [7]. To the best of our knowledge, no other work on FER has focused on this type of bias, and most have only focused on representational bias.

3 Methodology

3.1 Datasets

In this work, we employ the FER+ dataset[16], based on FER2013 [16]. FER2013 [16] is one of the most popular publicly available *in the wild* FER datasets, with more than $32,000$ labeled images obtained from Internet searches. The images in the original dataset were automatically annotated, leading to systematic inaccuracies, which were later corrected by FER+ [6], a relabeling of the same image set. The images in FER+ are grayscale and have a small resolution of 48 × 48 pixels. This small image size supports fast and resource-light model training, one of the main reasons for its popularity.

3.2 Demographic Relabeling

As FER2013 is not gender labelled, we use an external model, FairFace [18] to obtain an apparent gender prediction for each image. The FairFace model was trained on the homonymous dataset, composed of $108,501$ images labeled for apparent gender, apparent race, and apparent age. In the original experiments, the model achieved an accuracy greater than 92% for gender recognition in FairFace and three other demographic datasets. The model is publicly available[1].

It is important to note that FairFace comes with some serious limitations. Although this is particularly evident in the race categories, limited to six stereotypical groups, namely White, Black, East Asian, Southeast Asian, Latino, Indian, and Middle Eastern, it is also present in the gender category. For the creation of FairFace, as is still common for most gender-labeled datasets, external annotators manually labeled gender into a binary classification of *Male* and *Female*. This classification correlates with how many societies identify gender, but can easily misrepresent people, as is the case for binary and non-binary transgender people and other gender non-conforming individuals [19]. Nevertheless, as almost no datasets have the required demographic information, proxy labels such as the ones provided by FairFace give us a reasonable overview of the population of the datasets, even if they could be unreliable for the individual subjects. Additionally, as the real demographic information is also unknown to the trained FER models, if bias is present in them it must be based only on the physical appearance. Thus, we perform our analysis on these labels, as they can help uncover biases based on these apparent demographic characteristics, even if they do not always correlate with the true self-reported characteristics. Any bias based on the apparent characteristics predicted by the auxiliary model must be considered under these limitations, and further work must be done to test if the bias is still present when we consider the real demographic characteristics.

3.3 Generation of derivative datasets

To study the impact of stereotypical bias, we generate three types of datasets, namely, stratified, balanced, and biased. All of these are created as subsets from the original FER+ dataset.

Stratified subsets. To enable the comparison between different datasets, we implement a method to generate stratified subsets from a source dataset with a given target size, expressed as a ratio $r \in [0, 1]$ of the number of examples in the original dataset. To generate a stratified version, we consider both the

[1] https://github.com/joojs/fairface

set of target classification labels $L = \{\text{angry}, \text{disgust}, \text{fear}, \text{happy}, \text{sad}, \text{surprise}, \text{neutral}\}$, and the demographic groups of interest, in our case $S = \{\text{male}, \text{female}\}$ and their combinations defined by the Cartesian product $L \times S$. For each of these combinations independently, we perform a random subsample with target ratio r. This process guarantees that the relative proportions between each label in L, the demographic group in S, or the combination of both in $L \times S$ are kept, while the overall size is reduced by the desired ratio r (plus or minus some rounding error). Thus, the stratified datasets maintain the same stereotypical deviations and general demographic proportions as the source data set.

Balanced subsets. As the original FER+ dataset already contains some stereotypical bias [13], we generate a balanced version of it to serve as a general baseline. This dataset has the same proportions of each label in L as the original FER+, but for each of them, the proportions of the demographic groups in S are equalized. To generate this balanced dataset, we first calculate the most underrepresented group $(l, s) \in L \times S$ by calculating the imbalance ratio of each one in their respective label l:

$$\text{imb}(l, s) = \frac{|\{x | x \in D_l \,\text{and}\, x \in D_s\}|}{|\{x | x \in D_l\}|} \ , \tag{1}$$

where D_l denotes the subset of the dataset samples labeled with l and D_s the subset identified as part of the demographic group s.

After this, we subsample each of the groups independently according to:

$$\text{ratio}(l, s) = \frac{\min_{l' \in L, s' \in S} \text{imb}(l', s')}{\text{imb}(l, s)} \ . \tag{2}$$

The resulting dataset keeps the distribution of the target labels while making the demographic groups in each label and in the whole dataset equally represented.

Biased subsets. Finally, we also generate intentionally stereotypically biased datasets. These datasets are built from the balanced datasets, but inducing a certain amount of bias into one of the labels l with respect to a target demographic group s. The amount of induced bias $b \in [-1, 1]$ is applied as:

- If $b < 0$, a negative bias is introduced, that is, the target demographic group $\{x | x \in D_l \,\text{and}\, x \in D_s\}$ is reduced by the ratio $1 + b$. The examples labeled as l belonging to the other demographic groups are kept intact.
- If $b > 0$, a positive bias is introduced, that is, the target demographic group is left intact, reducing the representation of the rest of the samples $\{x | x \in D_l \,\text{and}\, x \notin D_s\}$ by the ratio $1 - b$.
- If $b = 0$, the balanced dataset is not modified, and no bias is introduced.

After biasing the target label l, the resulting number of examples of that label is $1 - \frac{|b|}{2}$ of the original label support. To compensate for this effect, the other labels are also subsampled by the ratio $1 - \frac{|b|}{2}$. This reduces the final dataset size, but keeps the label distribution equal to the original dataset.

The resulting dataset has, for $b = -1$ a total absence of the target demographic group in the label (underrepresentation), for $b = 1$ only samples of the target demographic group in the label (overrepresentation), and for $b = 0$ is balanced. The intermediate values allow for fine control of the amount of bias. In all cases, the label distribution is kept identical to the original dataset.

3.4 Experiments

In our experiments, we aim to generate biased datasets in the extremes of the stereotypical bias possibilities and then measure the final model accuracy imbalances for the relevant demographic groups and labels. For this, we first obtain the demographic profile of FER+ in the gender category. With this information, we chose some of the more heavily biased labels and generate datasets that exaggerate those same biases. The biased datasets are generated with different degrees of bias, from a negative bias of -1 to a positive one of 1 in steps of 0.2, all of them with respect to the ”female” class as recognized by FairFace. The balanced datasets will serve as a baseline, showing the behavior expected in the absence of stereotypical bias for a certain dataset size.

To analyze the influence of the datasets on the performance of the model, we train a model for each generated dataset and obtain the predictions over the whole FER+ test partition. We then obtain the recall for each combination of dataset, label, and gender group, that is, the accuracy of the classifier for the examples belonging to the specific gender group and with a certain true label. In particular, we expect to obtain the maximum difference in recall between the demographic categories *male* and *female* in the extreme biased datasets for the biased labels, as a measure of the maximum impact of stereotypical bias on the recognition of the affected labels.

3.5 Experimental Setup

We employ a simple VGG11 [27] network with no pretraining as the base test model. This is a classical convolutional architecture often used as a baseline for machine learning applications. The experiments are developed on PyTorch 1.10.0 and Fastai 2.6.3. The hardware used is a machine equipped with a GeForce RTX 2060 Super GPU, 20 GB of RAM, an Intel® Xeon® i5-8500 CPU, and running Ubuntu Linux 20.04.

All the models are trained under the same conditions and hyperparameters, namely, a maximum learning rate of $1e^{-2}$ with a 1cycle policy (as described in [28] and implemented in Fastai) for 20 iterations. This parameter was decided using the *lr_finder* tool in Fastai. The batch size is set to 256, the maximum allowed by the hardware setup. For each dataset, we train the model 10 times and average the results over them. We have also applied the basic data augmentation provided by Fastai through the *aug_transforms* method, including left-right flipping, warping, rotation, zoom, brightness, and contrast alterations.

For each dataset configuration to be tested, we perform ten individual training processes, for each one regenerating a new resampled dataset to ensure that the sampling process does not affect the final results.

4 Results and Discussion

4.1 Dataset initial bias

We perform the demographic relabeling of FER+ with the FairFace public model, as described in Section 3.2. The proportions of the gender category in the whole dataset and for each label are shown in Figure 1, together with the label supports. The global gender proportions are almost uniform, at 50.1% for the *Female* group and 49.9% for the *Male* group, showing very little direct representational bias. For stereotypical bias, the individual labels show a much greater disparity. The two extremes are the label *angry*, with an underrepresentation of the *Female* group (36.27% of the label support) and the label *happy*, with an underrepresentation of the *Male* group (38.7% of the label support). The rest of the labels in the dataset lie in between, with slightly lower imbalances.

Interestingly, the biases found in the labels *happy* and *angry* are consistent with the classical *angry-men-happy-women* bias, a psychological bias pattern well researched in the expression and recognition of human emotions [21,4].

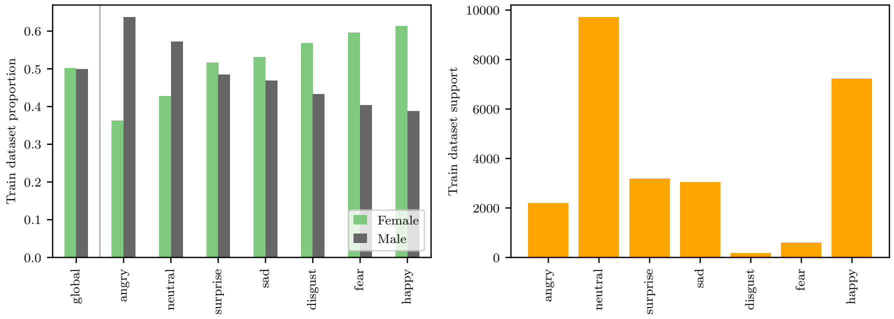

Fig. 1. FER+ gender distribution and support by label.

4.2 Induced bias impact

The recall results obtained by the models are shown in Table 1. For brevity, only the results for the four most extremely biased datasets and the size-equivalent stratified balanced dataset are reported in the Table, with the complete results being graphically presented in Figure 2. The difference between gender recalls is highest for the biased datasets in all cases, with the largest absolute difference found in the labels *angry*, *disgust*, and *happy*. In particular, for the *angry* label, biasing against the *Female* group maximizes the recall difference at 29.36% in favor of the *Male* group, while for the *happy* label it is the positive bias in favor of the *Female* group that maximizes the difference at 15.03%. Biasing the *angry* label generates a total range of disparity of 49.53% between its extremes, while biasing the *happy* label of 26.17%.

The label *disgust* seems to be a particular case, with the largest recall differences between the gender groups overall. Difference values range from a 8.33% difference to a 23.92% difference, always in favor of the *Male* group. Recall that none of the biased datasets are designed to bias in this label, that is kept balanced in all the derivative datasets, and even in the original dataset exhibits only a mild stereotypical bias against the *Male* group, which constitutes a 43.25% of the original support. However, this label also has the lowest support in the original FER+ dataset, with the lowest general recalls of all labels for all configurations. The label *disgust* also shows the highest standard deviation, between ±10.22 and ±16.83, making the results for this label unreliable.

The rest of the labels show some variations in general, but generally seem unaffected by the bias induced in *angry* and *happy*. An exception seems to be in the application of the positive bias in the *happy* label, overrepresenting the *female* group for that label, which seems to decrease the recall for the *angry* label of the same group.

		Happy			Angry		
		Female −1.00	Female 0.00	Female +1.00	Female −1.00	Female 0.00	Female +1.00
	Size	9475	9476	9475	9477	9476	9477
angry	Male	74.40 ± 1.47	76.30 ± 3.40	74.24 ± 4.73	**81.25 ± 2.61**	76.30 ± 3.40	55.87 ± 2.24
	Female	74.53 ± 3.45	73.21 ± 3.33	67.74 ± 3.34	51.89 ± 3.77	73.21 ± 3.33	**76.04 ± 1.99**
	Diff	0.13 ± 3.75	−3.10 ± 4.76	−6.50 ± 5.79	**−29.36 ± 4.59**	−3.10 ± 4.76	20.17 ± 2.99
neutral	Male	66.45 ± 3.10	71.52 ± 2.71	67.24 ± 4.66	70.72 ± 1.84	71.52 ± 2.71	**74.91 ± 1.66**
	Female	66.33 ± 2.77	67.29 ± 3.00	64.86 ± 3.79	67.99 ± 2.48	67.29 ± 3.00	**68.85 ± 1.26**
	Diff	−0.12 ± 4.15	−4.22 ± 4.04	−2.38 ± 6.01	−2.73 ± 3.09	−4.22 ± 4.04	**−6.06 ± 2.09**
surprise	Male	80.50 ± 3.91	83.71 ± 2.58	81.93 ± 3.16	82.77 ± 2.32	83.71 ± 2.58	**84.16 ± 1.71**
	Female	83.46 ± 3.06	84.88 ± 2.67	84.83 ± 2.51	**87.22 ± 2.75**	84.88 ± 2.67	86.29 ± 1.87
	Diff	2.97 ± 4.97	1.17 ± 3.71	2.90 ± 4.03	**4.45 ± 3.60**	1.17 ± 3.71	2.13 ± 2.54
sad	Male	62.22 ± 3.71	66.82 ± 1.53	66.88 ± 3.90	**69.66 ± 2.53**	66.82 ± 1.53	69.55 ± 2.69
	Female	67.88 ± 2.71	70.33 ± 2.30	68.91 ± 3.26	**74.29 ± 2.25**	70.33 ± 2.30	68.59 ± 2.72
	Diff	**5.66 ± 4.60**	3.51 ± 2.76	2.04 ± 5.09	4.63 ± 3.39	3.51 ± 2.76	−0.96 ± 3.82
disgust	Male	51.25 ± 16.01	45.00 ± 9.19	**61.25 ± 11.46**	56.88 ± 9.86	45.00 ± 9.19	56.25 ± 12.18
	Female	35.33 ± 5.21	36.67 ± 4.47	37.33 ± 6.11	**44.67 ± 7.33**	36.67 ± 4.47	40.00 ± 7.30
	Diff	−15.92 ± 16.83	−8.33 ± 10.22	**−23.92 ± 12.98**	−12.21 ± 12.29	−8.33 ± 10.22	−16.25 ± 14.20
fear	Male	66.25 ± 5.73	65.83 ± 4.86	**68.75 ± 4.66**	56.67 ± 7.73	65.83 ± 4.86	59.58 ± 7.23
	Female	59.76 ± 3.76	58.33 ± 4.16	57.14 ± 5.11	**60.95 ± 5.02**	58.33 ± 4.16	55.48 ± 6.21
	Diff	−6.49 ± 6.85	−7.50 ± 6.40	**−11.61 ± 6.91**	4.29 ± 9.21	−7.50 ± 6.40	−4.11 ± 9.53
happy	Male	89.10 ± 2.46	87.66 ± 2.23	72.91 ± 3.47	87.90 ± 2.11	87.66 ± 2.23	**89.79 ± 1.71**
	Female	77.96 ± 1.86	87.75 ± 2.14	87.94 ± 1.87	**88.42 ± 1.89**	87.75 ± 2.14	88.02 ± 1.48
	Diff	−11.14 ± 3.09	0.09 ± 3.09	**15.03 ± 3.94**	0.52 ± 2.83	0.09 ± 3.09	−1.77 ± 2.26

Table 1. Recall by label and gender for the key datasets analyzed. For the difference between gender recalls, the highest absolute value is in bold.

The difference between the recalls of the *male* and *female* groups for each degree of induced bias is shown in Figure 2. In the Figure, the vertical axis corresponds to the difference in recall from the *female* group to the *male* group, and the horizontal axis corresponds to the amount of induced bias. The difference of recall obtained in the balanced datasets is included as the comparison baseline.

For all labels, if no bias is introduced on that particular label, the recall differences are close to the baseline levels. When observing the differences in the recall of the affected label for the biased datasets, the effect of the dataset bias becomes apparent. For both the *angry* and *happy* labels, the negative biases, which correspond to an under-representation of the *female* group on the label, show a difference in recall in favor of the *male* group, and the opposite is observed for positive amounts of bias. For the datasets biased in the *angry* label with a negative or positive bias for the *female* group, the difference in recalls of the *angry* label quickly deviates from baseline levels when exposed to a bias of ±0.2 or greater, exceeding ±20% of difference under extreme bias conditions. For the label *happy*, the effect is not as pronounced and only when trained on datasets with bias of ±0.4 or higher does the difference in recalls deviate from the baseline behavior.

For both the labels *angry* and *happy*, a safe zone can be observed where bias in the dataset does not significantly affect the difference in recalls. The behavior of the biased model when trained under this limited amount of bias seems to be similar to the baseline dataset. In the case of the label *happy*, this safe zone includes the datasets with a stereotypical bias of ±0.4 and lower, while on *angry* it is more restricted, including only those with a stereotypical bias of ±0.2 and lower.

5 Conclusion

In this work, we have studied the impact of stereotypical bias in FER datasets and their resulting models through the induction of controlled bias in the dataset. In particular, for the FER problem, we have observed up to a 29% disparity in the recognition of certain emotions, namely *angry*, when the dataset lacks representation of a gender category for the label. We have shown that this kind of bias is already present in publicly available datasets, in particular in FER+, but our experiments suggest that a small amount of stereotypical bias in the gender category seems acceptable, not impacting the final performance for the under-represented group. Nevertheless, it is important to notice that the acceptable amount of stereotypical bias seems to be context-dependent, varying at least between labels. Our findings support the importance of a thorough bias analysis of public datasets in problems like FER, where a global balance of demographic representation in the dataset can still hide other types of bias that harm certain demographic groups.

In light of our findings, we highly recommend that future datasets, especially those created from Internet searches and intended for public release, are tested for stereotypical bias and corrected accordingly by down-sampling the overrep-

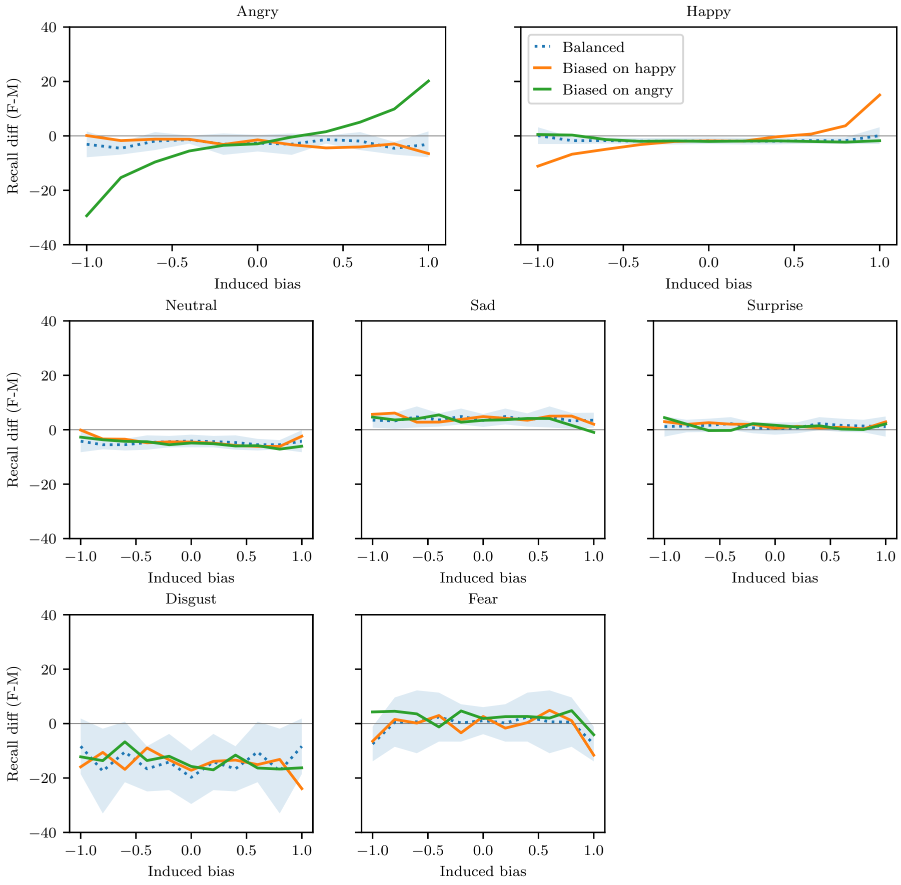

Fig. 2. Recall difference Male-Female in the different emotion labels. Positive numbers mean a higher recall for the *Female* group than for the *Male* one. The baseline balanced datasets are plotted according to size, aligned with the corresponding biased datasets.

resented demographic groups. Although other mitigation techniques could be performed later in the training phases, this type of bias is easy to overlook and can leak into bias in the trained models if left untreated. Furthermore, we strongly advise dataset creators to include the relevant demographic information of the subjects when possible, to allow the future study of new forms of demographic bias in their datasets.

A problem that requires further analysis is the large differences in the gender recall of certain labels, such as *disgust*. This difference is present even for the balanced versions of the dataset, suggesting a measurement bias or an inherent representation problem in this label. The label *disgust*, in particular, has low support, which could imply that stereotypical bias problems have a greater impact in smaller datasets. Further work is also required to replicate these results for other datasets, models and different applications. The development of properly labeled datasets that include demographic information of the represented subjects would also solidify this analysis, currently limited by the demographic relabeling model employed.

Acknowledgments

This work was funded by a predoctoral fellowship of the Research Service of Universidad Publica de Navarra, the Spanish MICIN (PID2019-108392GB-I00 and PID2020-118014RB-I00 / AEI / 10.13039/501100011033), and the Government of Navarre (0011-1411-2020-000079 - Emotional Films).

References

1. Abbasi, T.M., Friedler, S.A., Scheidegger, C., Venkatasubramanian, S.: Fairness in representation: Quantifying stereotyping as a representational harm. Proceedings of the 2019 SIAM International Conference on Data Mining (SDM) pp. 801–809 (May 2019). https://doi.org/10.1137/1.9781611975673
2. Ahmad, K., Wang, S., Vogel, C., Jain, P., O'Neill, O., Sufi, B.H.: Comparing the Performance of Facial Emotion Recognition Systems on Real-Life Videos: Gender, Ethnicity and Age. In: Arai, K. (ed.) Proceedings of the Future Technologies Conference (FTC) 2021, Volume 1, vol. 358, pp. 193–210. Springer International Publishing, Cham (2022). https://doi.org/10.1007/978-3-030-89906-6_14
3. Assuncao, G., Patrao, B., Castelo-Branco, M., Menezes, P.: An Overview of Emotion in Artificial Intelligence. IEEE Transactions on Artificial Intelligence pp. 1–1 (2022). https://doi.org/10.1109/TAI.2022.3159614
4. Atkinson, A.P., Tipples, J., Burt, D.M., Young, A.W.: Asymmetric interference between sex and emotion in face perception. Perception & Psychophysics **67**(7), 1199–1213 (Oct 2005). https://doi.org/10.3758/BF03193553
5. Avella, M.d.P.R.: Crime Prediction Artificial Intelligence and the Impact on Human Rights. Telecommunications System & Management **0**(0), 2–2 (Aug 2020)
6. Barsoum, E., Zhang, C., Ferrer, C.C., Zhang, Z.: Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution. arXiv:1608.01041 [cs] (Sep 2016)

7. Berk, R., Heidari, H., Jabbari, S., Kearns, M., Roth, A.: Fairness in Criminal Justice Risk Assessments: The State of the Art. Sociological Methods & Research **50**(1), 3–44 (Jul 2018). https://doi.org/10.1177/0049124118782533
8. Birhane, A., Prabhu, V.U., Kahembwe, E.: Multimodal datasets: Misogyny, pornography, and malignant stereotypes (Oct 2021)
9. Bordalo, P., Coffman, K., Gennaioli, N., Shleifer, A.: Stereotypes*. The Quarterly Journal of Economics **131**(4), 1753–1794 (Nov 2016). https://doi.org/10.1093/qje/qjw029
10. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Friedler, S.A., Wilson, C. (eds.) Proceedings of the 1st Conference on Fairness, Accountability and Transparency. Proceedings of Machine Learning Research, vol. 81, pp. 77–91. PMLR (Feb 2018)
11. Denton, E., Hanna, A., Amironesei, R., Smart, A., Nicole, H.: On the genealogy of machine learning datasets: A critical history of ImageNet. Big Data & Society **8**(2), 205395172110359 (Jul 2021). https://doi.org/10.1177/20539517211035955
12. Deuschel, J., Finzel, B., Rieger, I.: Uncovering the Bias in Facial Expressions. arXiv:2011.11311 [cs] (Nov 2021). https://doi.org/10.20378/irb-50304
13. Dominguez-Catena, I., Paternain, D., Galar, M.: Assessing Demographic Bias Transfer from Dataset to Model: A Case Study in Facial Expression Recognition (May 2022). https://doi.org/10.48550/arXiv.2205.10049
14. Domnich, A., Anbarjafari, G.: Responsible AI: Gender bias assessment in emotion recognition. arXiv:2103.11436 [cs] (Mar 2021)
15. Ekman, P., Friesen, W.V.: Constants across cultures in the face and emotion. Journal of Personality and Social Psychology **17**(2), 124–129 (1971). https://doi.org/10.1037/h0030377
16. Goodfellow, I.J., Erhan, D., Carrier, P.L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.H., Zhou, Y., Ramaiah, C., Feng, F., Li, R., Wang, X., Athanasakis, D., Shawe-Taylor, J., Milakov, M., Park, J., Ionescu, R., Popescu, M., Grozea, C., Bergstra, J., Xie, J., Romaszko, L., Xu, B., Chuang, Z., Bengio, Y.: Challenges in Representation Learning: A report on three machine learning contests. arXiv:1307.0414 [cs, stat] (Jul 2013)
17. Jannat, S.R., Canavan, S.: Expression Recognition Across Age. In: 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021). pp. 1–5 (Dec 2021). https://doi.org/10.1109/FG52635.2021.9667062
18. Karkkainen, K., Joo, J.: FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. In: 2021 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1547–1557. IEEE, Waikoloa, HI, USA (Jan 2021). https://doi.org/10.1109/WACV48630.2021.00159
19. Keyes, O.: The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. Proceedings of the ACM on Human-Computer Interaction **2**(CSCW), 1–22 (Nov 2018). https://doi.org/10.1145/3274357
20. Kim, E., Bryant, D., Srikanth, D., Howard, A.: 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, pp. 638–644. Association for Computing Machinery, New York, NY, USA (Jul 2021)
21. Kring, A.M., Gordon, A.H.: Sex Differences in Emotion: Expression, Experience, and Physiology. Journal of Personality and Social Psychology p. 18 (1998)
22. Li, S., Deng, W.: Deep Facial Expression Recognition: A Survey. IEEE Transactions on Affective Computing pp. 1–1 (2020). https://doi.org/10/gkk8dv

23. Mehrabian, A.: Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in Temperament. Current Psychology **14**(4), 261–292 (Dec 1996). https://doi.org/10.1007/BF02686918
24. Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M.E., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., Kompatsiaris, I., Kinder-Kurlanda, K., Wagner, C., Karimi, F., Fernandez, M., Alani, H., Berendt, B., Kruegel, T., Heinze, C., Broelemann, K., Kasneci, G., Tiropanis, T., Staab, S.: Bias in Data-driven AI Systems – An Introductory Survey. arXiv:2001.09762 [cs] (Jan 2020)
25. Prabhu, V.U., Birhane, A.: Large image datasets: A pyrrhic win for computer vision? arXiv:2006.16923 [cs, stat] (Jul 2020)
26. Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., Hall, P.: Towards a Standard for Identifying and Managing Bias in Artificial Intelligence. Tech. rep., National Institute of Standards and Technology (Mar 2022). https://doi.org/10.6028/NIST.SP.1270
27. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition (Apr 2015)
28. Smith, L.N.: A disciplined approach to neural network hyper-parameters: Part 1 – learning rate, batch size, momentum, and weight decay. arXiv:1803.09820 [cs, stat] (Apr 2018)
29. Suresh, H., Guttag, J.V.: A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. arXiv:1901.10002 [cs, stat] (Jun 2021)
30. Verma, S., Rubin, J.: Fairness definitions explained. In: Proceedings of the International Workshop on Software Fairness. pp. 1–7. ACM, Gothenburg Sweden (May 2018). https://doi.org/10.1145/3194770.3194776
31. Wang, T., Zhao, J., Yatskar, M., Chang, K.W., Ordonez, V.: Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations (Oct 2019)
32. Xu, T., White, J., Kalkan, S., Gunes, H.: Investigating Bias and Fairness in Facial Expression Recognition. In: Bartoli, A., Fusiello, A. (eds.) Computer Vision – ECCV 2020 Workshops. pp. 506–523. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-65414-6_35
33. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.W.: Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 2979–2989. Association for Computational Linguistics, Copenhagen, Denmark (2017). https://doi.org/10.18653/v1/D17-1323