Bias in Face Image Classification Machine Learning Models: The Impact of Annotator's Gender and Race

Andreas Kafkalias[1], Stylianos Herodotou[1], Zenonas Theodosiou[1], and Andreas Lanitis[1,2](✉)

[1] CYENS Center of Excellence, Nicosia, Cyprus
z.Theodosiou@cyens.org.cy

[2] Visual Media Computing Lab, Department of Multimedia and Graphic Arts,
Cyprus University of Technology, Limassol, Cyprus
andreas.lanitis@cut.ac.cy

Abstract. An important factor that ensures the correct operation of Machine Learning models is the quality of data used during the model training process. Quite often, training data is annotated by humans, and as a result, annotation bias may be introduced. In this study, we focus on face image classification and aim to quantify the effect of annotation bias introduced by different groups of annotators, allowing in that way the understanding of the problems that arise due to annotation bias. The results of the experiments indicate that the performance of Machine Learning models in several face image interpretation tasks is correlated to the self-reported demographic characteristics of the annotators. In particular, we found significant correlation to annotator race, while correlation to gender is less profound. Furthermore, experimental results show that it is possible to determine the group of annotators involved in the annotation process by considering the annotation data provided by previously unseen annotators. The results emphasize the risks of annotation bias in Machine Learning models.

Keywords: Machine learning · Annotation bias · Face images

1 Introduction

Over the last decade, the use of Machine Learning (ML) has increased dramatically [11] as numerous daily tasks are accomplished based on ML models. For example, ML has been used in recommendation systems, speech recognition, robot control, medical diagnosis, natural language processing, weather forecast, biometric authentication, text/image synthesis and for many other applications.

This project is funded by the Cyprus Research and Innovation Foundation under grant EXCELLENCE/0918/0086 (DESCANT) and by the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 739578 (RISE).

© IFIP International Federation for Information Processing 2022
Published by Springer Nature Switzerland AG 2022
I. Maglogiannis et al. (Eds.): AIAI 2022, IFIP AICT 647, pp. 89–100, 2022.
https://doi.org/10.1007/978-3-031-08337-2_8

At its core, an ML model will only be as good as the data used for training the model. The main issue that relates to the quality of a training dataset is how well training samples represent the classes to be classified, in terms of quality and quantity. Furthermore, an important aspect of the training data is the quality of the annotation, as imperfections in the annotation process can influence the training data quality. Quite often, the annotation process requires human expertise, and as a result it is subjected to the expression of social stereotypes. This is because as social beings, humans are continuously engaged in a process of interpreting and forming impressions of others. However, cognitive heuristics often lead us to make trait inferences and evaluations of others that are based on very little concrete evidence (i.e., social stereotyping) [12]. For example, political candidates whose facial appearance is regarded as more accomplished, have a higher chance of winning the elections [20]. The process of data annotation can be influenced by social stereotyping and introduce bias in ML models, that eventually affects their performance.

In this study, we aim to quantify the effect of annotation bias, in terms of the performance of ML models, allowing in that way the understanding of the problems that arise due to the expression of social stereotypes in the annotation process. In particular, we compare the performance of ML models trained using data annotated by male annotators, female annotators, and annotators belonging to different (self-reported) racial groups. All groups of annotators annotated face images in relation to the classification tasks of gender recognition, race classification, attractiveness estimation, and trustworthiness estimation of subjects shown in face images. The comparison of the performance of ML models trained using data annotated by different annotator groups, allows the derivation of conclusions related to the effects of social bias (stereotyping) in ML. To further emphasize the extent of the annotation bias problem, we also present results that show that it is possible to determine the group of annotators involved in the annotation process, by considering the annotation data provided.

2 Background and Literature Review

Since the main classification tasks considered in this paper relate to face image interpretation, a brief review of the relevant literature of this topic is provided, followed by a review of the work related to bias in ML.

2.1 Face Image Interpretation

Zhao et al. [24] provide a thorough survey of the conventional methods used for face recognition where they present the main steps that include the tasks of face detection and feature extraction. They also elaborate on the methods used in the face recognition step which they divide into three categories: holistic matching methods, feature-based matching methods and hybrid methods. More recent surveys on face recognition focus on the use of deep networks by introducing different dedicated network architectures used for face image recognition [9,15].

Apart from the face recognition task, other surveys focus on different face image interpretation tasks such as emotion recognition [14], age estimation [18], and pose estimation [16].

In this paper, we focus on the tasks of gender recognition, race recognition, attractiveness estimation and trustworthiness estimation. While for the case of gender and race recognition, a plethora of techniques were reported in the literature [8,17], only few attempts were recorded for the problems of attractiveness estimation and trustworthiness estimation. Todorov et al. [21] build a model for representing face trustworthiness using a computer model for face representation. Using this model, they generated novel faces with an increased range of trustworthiness. Xu et al. [23] propose the use of the Hierarchical Multi-task Network (HMTNet) network, that performs gender, race and facial attractiveness estimation simultaneously. Experimental results reported for the combined gender, race and attractiveness estimation tasks, outperform the results obtained by other deep architectures.

2.2 Bias in Computer Vision Algorithms

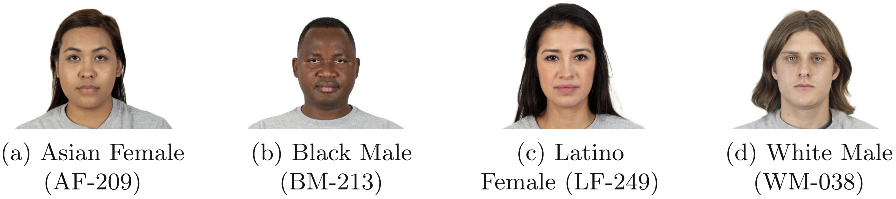
Fabbrizi et al. [6] present an overview of the major biases encountered in computer vision tasks that include selection, framing and label biases. Selection bias can be characterised as "any disparities or associations created as a result of the process by which subjects are included in a visual dataset" [6]. Selection bias is encountered in numerous ML models. For example Kay et al. [13] examined the representation of men and women in different occupations, in order to demonstrate that selection bias exists in search engines. More specifically, they were able to show that in male-dominated occupations, the male gender dominance is even more present in the Google's image search engine. However, for the respective female-dominated occupations, the results are more balanced. This showcases systematic selection biases in their retrieval algorithms and proves the importance of data gathering processes.

Framing biases in Computer Vision can arise by selecting or choosing specific characteristics and aspects in visual datasets which mislead and cause interpretation issues for the image portrayed. According to Coleman [5], this is usually achieved by manipulation of an image through editing, cropping or selecting a particular view/angle. The work of Heuer et al. [10] on the depiction of obesity in US online news websites demonstrate how framing biases can affect the meaning of a picture. Their analysis shows how obese people are portrayed with negative characteristics such as cropped heads, while non-obese people are portrayed without such characteristics. Such cruel effects will influence the opinion of the common viewer and therefore the meaning of the image.

Label bias is usually introduced during the annotation process, where different annotators may produce misleading labels. For example, social factors, such as the global pandemic of COVID-19 or the Black Lives Matter movement, or even personal circumstances may influence the annotation process. Christoforou et al. [4] focus on label bias and the limitations of crowdsourced data. They address this issue by clustering annotated data for face images collected before and after the COVID-19 pandemic. Using the Chicago Face Database they created two clusters based on health-related and identity tags. They demonstrated that temporal variations affect the annotation of data based on crowdsourcing. Christoforou et al. [4] points out the limitations that emerge with crowdsourced data based on the influence of consequential events around the globe, which is something that the requester must recognise, manage and ensure to raise awareness to the annotators. Torralba and Efros [22] suggest that label biases can also come up as different annotators can think up a variety of labels for a single object. This usually appears within enormous datasets where an object can take multiple names. Even though face classification is not very complex, the process of annotating a face dataset can be affected by the bias and viewpoint of the annotators. Previous attempts to create face datasets based on the opinion of annotators have shown that their opinion will reflect heavily on the labelling. For example, Liang et al. [7] showcased the bias of the annotators that participated in the study by creating a facial beauty dataset based on features such as attractiveness.

3 Face Database and Annotation

The data used for the project are images from the Chicago Face Database (CFD) [1] and more precisely, the main CFD image set that consists of 597 face images of unique individuals. The CFD includes images of men and women, belonging to four racial groups. Figure 1 shows typical samples of images from the CFD used in our experiments. All images in the CFD were annotated on average by 47 annotators and the mean among all annotations provided is considered as the ground truth, as is common practice in the field. Ground truth includes labels for the classification tasks considered in this work such as the gender of each subject (Male/Female), race (Latino, Asia, Black White, and additional mixed races), attractiveness (scale 1 to 7, where 1 means lowest attractiveness), and trustworthiness (scale 1 to 7, where 1 means lowest trustworthiness).



(a) Asian Female (AF-209)

(b) Black Male (BM-213)

(c) Latino Female (LF-249)

(d) White Male (WM-038)

Fig. 1. Sample of images used in modelling from the CDF. The tags in the brackets represent the gender, race and their ID in the Database.

For the needs of this work, a dedicated annotation process was set up through Clickworker [2]. Clickworker is an online platform where freelancers in their own free time, get paid for micro jobs such as image annotation. Annotators from different racial groups, different genders and different ages were invited to participate. During the annotation process, annotators had to specify the gender, race (Asian, Black, Latino, White, Multi-race or Other), level of attractiveness (scale 1 to 7), and the level of trustworthiness (scale 1 to 7) of each subject shown in an image. Furthermore, annotators were asked to provide information about their own gender, race, age, and employment status.

Three hundred eighty-eight annotators participated in the experiment. Among the annotators 52% identified themselves as males, 47% as females and 1% as other. Regarding their race, 69% identified themselves as White, 12% as Black, 9% as Asian, 6% as Latino, and 2% as Multi-racial and the rest 2% as other.

Every image in the CFD was labelled by at least four different annotators, resulting in a dataset of 2370 different entries. This implies that some of the images were not annotated by all different genders and races. The label of an image under any race or gender was chosen by the majority of the corresponding category. In the case of race, although annotators could indicate six different labels (Asian, Black, Latino, Mixed, Other or White), only the labels Asian, Black, Latino, and White were considered as only in very few occasions the labels Mixed and Other were indicated by annotators. Furthermore the ratings of trustworthiness and attractiveness provided were mainly in the range of 3 to 5, rather than receiving values covering the full 1 to 7 scale. For this reason, responses for the ratings of the two quantities were re-scaled to cover the whole range of 1 to 7. Given the uncertainly in providing an exact value for trustworthiness and attractiveness, the corresponding ratings were divided into the three categories of low, medium and high. Within this context data within the range of 1 to 3 was assigned to the low value, ratings of 4 was given the medium and ratings in the range of 5 to 7 were given a high value. As a result, the trustworthiness and attractiveness estimation problems were posed as three-class classification problems.

4 Experiment 1: Comparing the Performance of Annotator-Specific Classification Models

The aim of this experiment is to compare the performance of ML models trained using the collected data, as to quantify the extent of possible bias introduced by different groups of annotators.

4.1 Model Training

During the process of model training, nine different Deep Learning Models were trained for each of the four tasks of gender, race, attractiveness and trustworthiness classification. For each model trained, the training data used is the one provided by the six groups of annotators (Male, Female, Asian, Black, Latino, White). Furthermore, for each classification task a model was trained based on the ground truth provided with the Chicago Face Database, and an additional model was trained based on randomly annotated data. To compensate for the fact that the vast majority of the annotated samples were attributed to White annotators, a randomly selected subset of samples annotated by White annotators was selected, where the number of observations in that case was on par with the numbers of observations from Asian, Black and Latino annotators. The model trained using the subset of white annotators was called "Reduced White" (RWh).

Model training was done using the lobe.ai tool [3]. By using open-source Machine Learning Architectures, the lobe.ai tool is able to automate Deep Learning classification tasks without the need to perform a rigorous manual model optimisation process, ensuring that all models under comparison are trained using exactly the same training and model optimisation procedures. Furthermore, the lobe.ai tool is able to achieve an excellent performance at low computational costs. However, although lobe.ai allows the export of trained models for use in conjunction with the most popular deep learning libraries, it does not provide explicit details of the model architecture and/or the training algorithms used for training the models. After loading the data with the appropriate labels, the lobe.ai tool needs approximately around 10–15 min in training and optimizing the models when the model training procedure was run on an AMD Ryzen 3600 6-Core Processor with 16 GB RAM.

4.2 Results and Discussion

Details of all models trained in terms of the number of samples and the performance achieved on the train and test data is shown in Table 1.

Table 1. Models' train and test accuracy for each classification tasks divided by the respective annotation categories. [3]

Model	Num of samples	Gender		Race		Attractiveness		Trustworthiness	
		Train	Test	Train	Test	Train	Test	Train	Test
Ground Truth (GT)	597	99	97	93	91	77	76	62	60
All Annotators (AA)	597	99	98	86	80	62	53	56	36
Male (Ma)	596	92	94	72	32	66	53	59	39
Female (Fe)	592	93	95	84	76	68	50	56	38
Asian (As)	106	94	88	90	65	36	28	44	27
Black (Bl)	165	95	81	80	56	74	36	75	36
Latino (La)	114	89	86	75	76	49	23	70	34
White (Wh)	597	94	93	82	69	70	47	57	39
Reduced White (RWh)	128	92	88	79	65	79	45	62	31
Random (Ra)	597	67	37	54	17	47	27	62	36

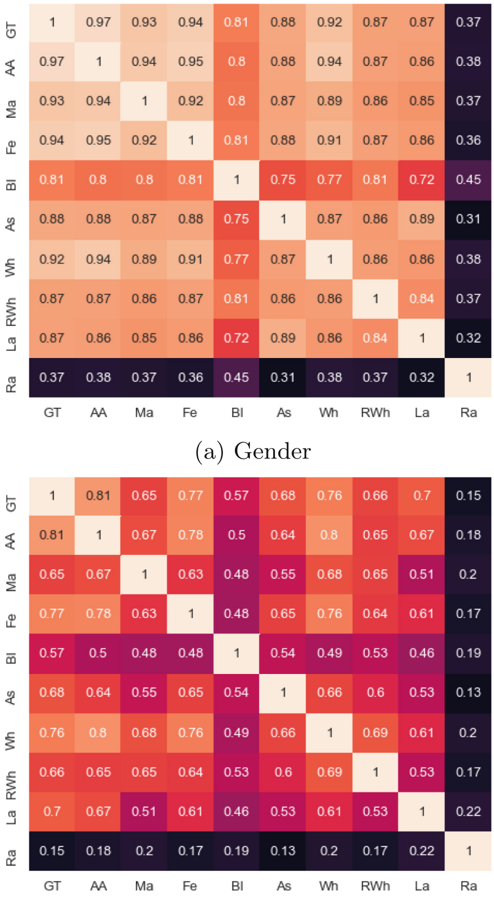
As expected models built using the ground truth data outperform the rest of the models. On the other hand, in most cases models built on random data have the worst performance. Apart from the race classification task, models trained using data annotated by male and female annotators have similar performance indicating that, for the tasks considered, the annotation process by male and female annotators leads to models with similar performance. However, models trained using data annotated by annotators belonging to different racial groups display increased diversity in performance.

With the introduction of deep learning and convolutional networks, tasks such as gender classification are now considered trivial for ML problems with expected accuracy of around 95%. However, models trained using data annotated by annotators from different racial backgrounds resulted in worse performance compared to the models built using data from annotators of different genders. Among the classification problems considered, the task of trustworthiness estimation received the lowest classification rates, implies that trust cannot be easily determined based only on facial appearance. For the attractiveness task, the models trained based on the Asian and Latino annotators achieve the worst performance, on par with the performance achieved by the models based on random annotations. This observation can be linked to different attitudes related to attractiveness cultivated in different cultures [19].

Comparing the results of the models built from the entire dataset of the White annotators (‘White’ model) against the models built with a randomly selected subset of images annotated by White annotators (‘Reduced White’ model), it is observed that although the models trained using reduced samples achieved lower performance, no major differences in relation to the comparison against models trained using Asian, Latino and Black narrators is observed. Therefore, we can conclude that the results related to the performance of the model trained using White annotators, are not attributed to the higher number of samples used during the training process.

The correlation matrices shown in Fig. 2 demonstrate the percentage agreement between the classification performance achieved by different models, for each task considered in the evaluation. For the tasks of gender and race classification (Fig. 2.(a) and Fig. 2.(b)) most of the models achieve high percentage agreement between each other, with an average of around 90% and 75% respectively. Excluding the models based on Black annotators and random models, the rest of the models have a similar agreement. Even though models based on Black annotators have a high percentage of test accuracy, it is clear that they disagree the most with the rest of the models with an average of 80% for gender and 50% for ethnicity. The contradiction between the agreement of models is attributed to label bias introduced during the annotation process. Based on data from Fig. 2, it is evident that Black annotators classify ethnicity and gender in a different way and this impacts the result of the respective models.

All models built for the trustworthiness estimation task, have very low agreement with each other. Furthermore, classification results for the task of trustworthiness estimation display high diversity among different groups of annotators, indicating the perception of those attributes varies from gender to gender and race to race. The results clearly demonstrate that trustworthiness estimation is subjected to bias due to the annotation process, hence a proper annotation procedure that involves annotators from different groups need to be employed to produce training data suitable for this challenging task. Except for the task of trustworthiness estimation, the models trained based on all white annotators and the models trained using the reduced subset of white annotators have high level of agreement. This indicates that when compared with the annotation bias introduced by annotators from different backgrounds, the bias introduced by different sizes of training samples is less important for the tasks of gender, race and attractiveness classification.



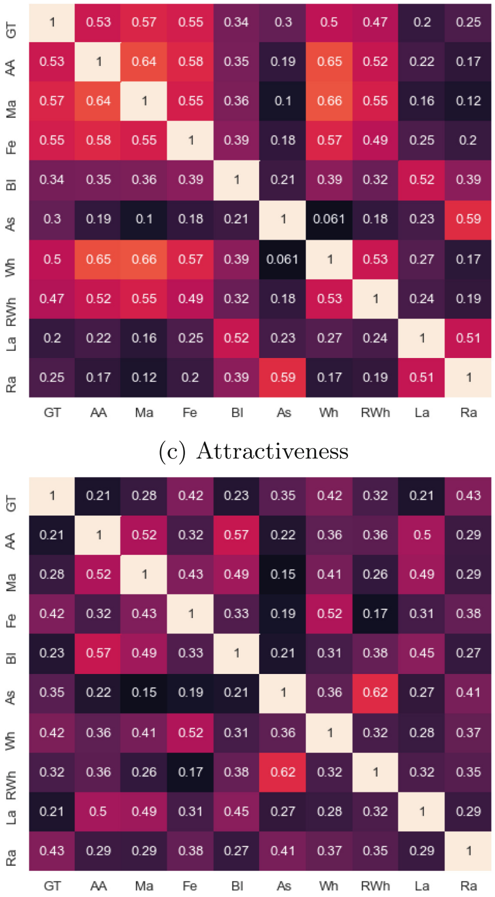
	GT	AA	Ma	Fe	Bl	As	Wh	RWh	La	Ra
GT	1	0.97	0.93	0.94	0.81	0.88	0.92	0.87	0.87	0.37
AA	0.97	1	0.94	0.95	0.8	0.88	0.94	0.87	0.86	0.38
Ma	0.93	0.94	1	0.92	0.8	0.87	0.89	0.86	0.85	0.37
Fe	0.94	0.95	0.92	1	0.81	0.88	0.91	0.87	0.86	0.36
Bl	0.81	0.8	0.8	0.81	1	0.75	0.77	0.81	0.72	0.45
As	0.88	0.88	0.87	0.88	0.75	1	0.87	0.86	0.89	0.31
Wh	0.92	0.94	0.89	0.91	0.77	0.87	1	0.86	0.86	0.38
RWh	0.87	0.87	0.86	0.87	0.81	0.86	0.86	1	0.84	0.37
La	0.87	0.86	0.85	0.86	0.72	0.89	0.86	0.84	1	0.32
Ra	0.37	0.38	0.37	0.36	0.45	0.31	0.38	0.37	0.32	1

(a) Gender

	GT	AA	Ma	Fe	Bl	As	Wh	RWh	La	Ra
GT	1	0.81	0.65	0.77	0.57	0.68	0.76	0.66	0.7	0.15
AA	0.81	1	0.67	0.78	0.5	0.64	0.8	0.65	0.67	0.18
Ma	0.65	0.67	1	0.63	0.48	0.55	0.68	0.65	0.51	0.2
Fe	0.77	0.78	0.63	1	0.48	0.65	0.76	0.64	0.61	0.17
Bl	0.57	0.5	0.48	0.48	1	0.54	0.49	0.53	0.46	0.19
As	0.68	0.64	0.55	0.65	0.54	1	0.66	0.6	0.53	0.13
Wh	0.76	0.8	0.68	0.76	0.49	0.66	1	0.69	0.61	0.2
RWh	0.66	0.65	0.65	0.64	0.53	0.6	0.69	1	0.53	0.17
La	0.7	0.67	0.51	0.61	0.46	0.53	0.61	0.53	1	0.22
Ra	0.15	0.18	0.2	0.17	0.19	0.13	0.2	0.17	0.22	1

(b) Ethnicity

Fig. 2. Correlation Matrices between the models of each category. Light colour indicates high percentage agreement while darker colours low percentage agreement.



(c) Attractiveness

(d) Trustworthiness

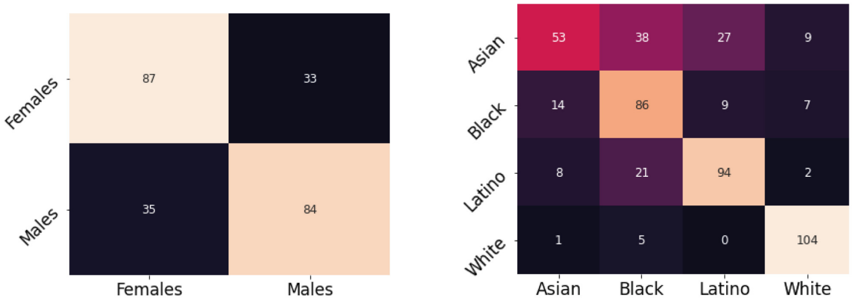
Fig. 2. (*continued*)

5 Experiment 2: Predicting Annotator Groups Based on Annotations

The results of experiment 1 show clear differences in performance, and disagreement between models trained using data annotated by different groups of annotators. To further investigate this phenomenon, the possibility of predicting the gender and the race of an annotator, based on their respective annotations was examined. In this context two classification models were trained. Each of those models take as input the ground truth values of gender, race, attractiveness and trustworthiness for each sample, along with the annotation provided by each annotator, using the data collected through the clickworkers platform (see Sect. 3). The models trained use a Multilayer Perceptron (MLP) architecture with eight inputs, four fully connected layers of 128 neurons with relu activation, a fully connected layer with 64 neurons and relu activation, and an output fully connected layer with a sigmoid activation. The outputs of the model corresponds to the gender and race of an annotator.

Once trained, the models are able to identify correctly the gender with a test accuracy of 70% for the ethnicity and 66% for the gender. The confusion matrices below in Fig. 3 demonstrate the results for the prediction of the annotators.

A cross validation with K folds, where K = 30, was run, in combination with a hypothesis z-test, to examine if the trained models can be used to predict the attributes of annotators with better accuracy than random guesses. The results indicate that there is a statistically significant improvement between the predicted labels of the trained models when compared to random guesses, for both the gender ($z = 5.57 \mid a = 0.01$) and race ($z = 20.54 \mid a = 0.01$). Based on these results, it is evident that the bias in the annotations is reflected as bias in the models predictions, and using these predictions it is possible to reverse engineer the process and identify attributes of an annotator.



(e) Gender

(f) Ethnicity

Fig. 3. Confusion matrices for both models created for predicting the features of the annotators. Light colour indicates a high prediction count while darker colours a lower prediction count.

6 Conclusions and Future Work

Experimental results presented in this paper demonstrate that the perception of characteristics such as attractiveness and trustworthiness vary as a function of annotator demographics (gender and race). In fact, sometimes machine learning models trained to classify those attributes have higher levels of agreements with models based on random data instead of another category of annotators. Even binary gender classification, which can be considered trivial in face interpretation, show different results when models are trained on data from annotators from different racial backgrounds. Furthermore, the bias in the models predictions makes it possible to reverse engineer and identify attributes of an annotator.

Based on the results obtained, it is evident that computer vision tasks that rely on training data annotated by humans could be heavily influenced by social stereotyping, that can cause biased performance. The work presented in this study provides quantitative results indicating the extend of the problem in several classification tasks, against the groups of annotators used, providing in that way useful insight for researchers involved in similar classification tasks.

The results of this study are demonstrated through an interactive tool at http://descant.cyens.org.cy/, so that the results of this project can be used by Machine Learning practitioners and students, as training material to anticipate the dangers of annotation bias.

In the future, we plan to extend our work to additional classification tasks, and test the extend of stereotype thread bias in different network architectures. Furthermore, we plan to use the lessons learned as part of this effort, to provide ways in which machine learning models can be trained to eliminate the effects of social bias, through a dedicated machine learning process.

Acknowledgements. This project is partially funded by the Cyprus Research and Innovation Foundation under grant EXCELLENCE/0918/0086 (DESCANT) and by the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 739578 (RISE). We would like to thank Dr. Jahna Otterbacher and Dr. Evgenia Christoforou for their contribution in formulating the experiment and for setting up the data collection process.

References

1. Chicago Face Database. https://www.chicagofaces.org/. Accessed 22 Feb 2022
2. Clickworker crowdsourcing. https://www.clickworker.com/. Accessed 22 Feb 2022
3. Lobe.ai webpage. https://www.lobe.ai/. Accessed 22 Feb 2022
4. Christoforou, E., Barlas, P., Otterbacher, J.: It's about time: a view of crowdsourced data before and during the pandemic. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1–14 (2021)
5. Coleman, R.: Framing the pictures in our heads: exploring the framing and agenda-setting effects of visual images. In: Doing News Framing Analysis, pp. 249–278. Routledge (2010)
6. Fabbrizzi, S., Papadopoulos, S., Ntoutsi, E., Kompatsiaris, I.: A survey on bias in visual datasets. arXiv preprint arXiv:2107.07919 (2021)
7. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. In: 2004 Conference on Computer Vision and Pattern Recognition Workshop, p. 178. IEEE (2004)
8. Fu, S., He, H., Hou, Z.G.: Learning race from face: a survey. IEEE Trans. Pattern Anal. Mach. Intell. **36**(12), 2483–2509 (2014)
9. Guo, G., Zhang, N.: A survey on deep learning based face recognition. Comput. Vis. Image Underst. **189**, 102805 (2019)
10. Heuer, C.A., McClure, K.J., Puhl, R.M.: Obesity stigma in online news: a visual content analysis. J. Health Commun. **16**(9), 976–987 (2011)
11. Jordan, M.I., Mitchell, T.M.: Machine learning: trends, perspectives, and prospects. Science **349**(6245), 255–260 (2015)

12. Jussim, L., Nelson, T.E., Manis, M., Soffin, S.: Prejudice, stereotypes, and labeling effects: sources of bias in person perception. J. Pers. Soc. Psychol. **68**(2), 228 (1995)
13. Kay, M., Matuszek, C., Munson, S.A.: Unequal representation and gender stereotypes in image search results for occupations. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, pp. 3819–3828 (2015)
14. Ko, B.C.: A brief review of facial emotion recognition based on visual information. Sensors **18**(2), 401 (2018)
15. Masi, I., Wu, Y., Hassner, T., Natarajan, P.: Deep face recognition: a survey. In: 2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), pp. 471–478. IEEE (2018)
16. Murphy-Chutorian, E., Trivedi, M.M.: Head pose estimation in computer vision: a survey. IEEE Trans. Pattern Anal. Mach. Intell. **31**(4), 607–626 (2008)
17. Ng, C.-B., Tay, Y.-H., Goi, B.-M.: A review of facial gender recognition. Pattern Anal. Appl. **18**(4), 739–755 (2015). https://doi.org/10.1007/s10044-015-0499-6
18. Panis, G., Lanitis, A.: An overview of research activities in facial age estimation using the FG-NET aging database. In: Agapito, L., Bronstein, M.M., Rother, C. (eds.) ECCV 2014. LNCS, vol. 8926, pp. 737–750. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16181-5_56
19. Rhodes, G., et al.: Attractiveness of own-race, other-race, and mixed-race faces. Perception **34**(3), 319–340 (2005)
20. Said, C.P., Sebe, N., Todorov, A.: Structural resemblance to emotional expressions predicts evaluation of emotionally neutral faces. Emotion **9**(2), 260 (2009)
21. Todorov, A., Baron, S.G., Oosterhof, N.N.: Evaluating face trustworthiness: a model based approach. Soc. Cogn. Affect. Neurosci. **3**(2), 119–127 (2008)
22. Torralba, A., Efros, A.A.: Unbiased look at dataset bias. In: CVPR 2011, pp. 1521–1528. IEEE (2011)
23. Xu, L., Fan, H., Xiang, J.: Hierarchical multi-task network for race, gender and facial attractiveness recognition. In: 2019 IEEE International Conference on Image Processing (ICIP), pp. 3861–3865. IEEE (2019)
24. Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A.: Face recognition: a literature survey. ACM Comput. Surv. (CSUR) **35**(4), 399–458 (2003)