AESTHETICNET: REDUCING BIAS IN FACIAL DATA SETS UNDER ETHICAL CONSIDERATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Facial Beauty Prediction (FBP) aims to develop a machine that can automatically evaluate facial attractiveness. Usually, these results were highly correlated with human ratings, and therefore also reflected human bias in annotations. Everyone will have biases that are usually subconscious and not easy to notice. Unconscious bias deserves more attention than explicit discrimination. It affects moral judgement and can evade moral responsibility, and we cannot eliminate it completely. A new challenge for scientists is to provide training data and AI algorithms that can withstand distorted information. Our experiments prove that human aesthetic judgements are usually biased. In this work, we introduce AestheticNet, the most advanced attractiveness prediction network, with a Pearson correlation coefficient of 0.9601, which is significantly better than the competition. This network is then used to enrich the training data with synthetic images in order to overwrite the ground truth values with fair assessments.

We propose a new method to generate an unbiased CNN to improve the fairness of machine learning. Prediction and recommender systems based on Artificial Intelligence (AI) technology are widely used in various sectors of industry, such as intelligent recruitment, security, etc. Therefore, their fairness is very important. Our research provides a practical example of how to build a fair and trustable AI.

1 MOTIVATION

In 2016 *Beauty.AI*, a Hong-Kong based technology company, hosted the first international beauty contest judged by artificial intelligence (beauty.ai, 2016) but the results were heavily biased, for example, against dark skin (Levin, 2016) subjects. "Machine learning models are prone to biased decisions, due to biases in data-sets" (Sharma et al., 2020). Biased training data potentially leads to discriminatory models, as the datasets are created by humans or derived from human activities in the past, for example hiring algorithms (Bogen, 2019). The reason for racist and discriminatory tendencies must be identified. As the learning algorithms become more complex, understanding why the decisions are made, or even how, prove to be nearly impossible (Bostrom & Yudkowsky, 2018). Therefore, the development of non-biased and fair training data and AI algorithms (defined by the European Commission High-Level Expert Group on Artificial Intelligence (European Commission High-Level Expert Group on Artificial Intelligence [AI HLEG], 2019)) is a new and increasingly complex challenge for scientists around the world (Bellamy et al., 2018). The specific field of aesthetic judgement is especially vulnerable to being biased, as aesthetic judgement itself is already a subjective rating (Richmond, 2017).

The purpose of facial beauty prediction (FBP) research is to classify images mimicking subjective human judgements. Investigations related to machine perception in a ground-truth free setting show that the data source depends on the measurement of human perception (Prijatelj et al., 2020). Therefore, artificial networks need a process to determine labels of the average person's judgement. Our data analysis has already proven that people consider their own ethnicity to be more attractive than others (Gerlach et al., 2020), this is the major bias in our experiments and within our dataset. With this tendency, it becomes difficult to generate input data to train a machine-learning algorithm, which assesses a person's attractiveness without bias.

This work not only helps to achieve moral enhancement through AI (see appendix B.1), but also helps eliminating social problems with this new technology (see appendix B.2).

2 STATE OF THE ART

While research on the estimation of images or portraits is not a new trend, it has gained increasing attention since the emergence of artificial intelligence (Zhang & Kreiman, 2021). Although, for many applications like autonomous driving, or image classification, AI undoubtedly is the best solution, applications that are affected by unconscious bias, like beauty prediction (Dornaika et al., 2020), tend to reflect bias that is likely to be prevalent within given datasets. Especially, when people subjective preferences play a role, such as in attractiveness judgement (Shank & DeSanti, 2018) or human resource evaluation (Lloyd, 2018), bias is almost certain to happen. Carrera (2020) conducted a piece of research on the implication of racism in image databases, that analysed the association of aggressiveness, kindness, beauty and ugliness with different images and found that the decisions of many people are affected by subconscious racism. Since researchers are aware of such effect, they found different ways to reduce subconscious bias in machine learning. Since the problem originates from the given databases, either the databases, or the training need to be changed.

The possibilities to change the databases include adding data, also referred to as *fair pre-processing* (Bellamy et al., 2018), either by selection or augmentation to insert underrepresented samples. While, deleting images is usually a bad idea, since it increases the chances of the network overfitting, it could theoretically be used to eliminate overrepresented images. On the other side, training can be altered by selecting only images, that do not increase the variance of each class currently used as training input. For example, variational autoencoders can be used to extract the features of the image, to later determine their variance, and only select images as input that do not increase the variance within given classes. Bellamy et al. (2018) also describe a third method, they called *fair post-processing*. Since their pipeline aimed to create debiased databases, the post-processing step is usually not applicable for most machine learning applications not creating databases.

3 BIASED AI

3.1 BIAS FROM HUMAN INDICATIONS

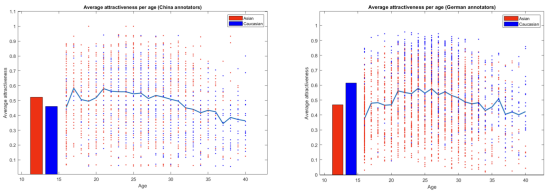

Figure 1: Chinese and European annotations, the red bar represents the score of Asian, the blue bar represents the score of European faces.

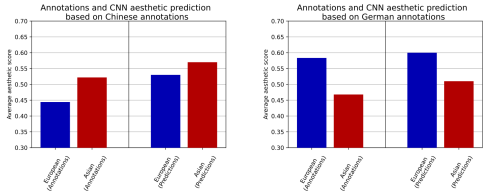

Figure 2: AestheticNet is trained on German or Chinese annotations only. The trained network follows the bias from the annotations.

First, we propose hypothesis 1: The results of the evaluation of the attractiveness of female pictures in the Asia-Europe data set by annotators in China and Germany are implicitly biased. We use our latest data set, which includes a total of 12,034 images of people from different social and ethnical backgrounds, with a total of 5.4 million annotations. Chinese and German participants have rated the pictures in the data set. We then have a comparison result to prove whether hypothesis 1 is true and mark this evaluation result as a ground truth. Figure 1 and fig. 2 confirm the statement of the first hypothesis. Further more, in this process, the results of our research also shows that aesthetic bias is not only related to ethnic background, but also related to age which also has been proven by other researchers (Gerlach et al., 2020), (Akbari et al., 2020).

3.2 AI TAKES ON HUMAN BIAS

We propose hypothesis 2: artificial intelligence will copy the human bias. We use convolutional neural networks (CNN) to predict facial aesthetic scores and introduce AestheticNet.

Related Work. With the introduction of CNNs and large-scale image repositories, facial image and video tasks get more powerful (Krizhevsky et al., 2017; Zeiler & Fergus, 2013; Deng et al.,

2009). Xie et al. (Xie et al., 2015a) present the SCUT-FBP500 dataset, containing 500 Asian female subjects with attractiveness ratings. Since "FBP is a multi-paradigm computation problem" the successor SCUT-FBP5500 (Liang et al., 2018) is introduced in 2018, including an increased database of 5500 frontal faces with multiple attributes: male/female, Asian/Caucasian, age, beauty score. Liang et al. (2018) have evaluated their database "using different combinations of feature and predictor, and various deep learning methods" on AlexNet (Krizhevsky et al., 2017), ResNet-18 (He et al., 2015) and ResNeXt-50 and achieved the Pearson Correlation *PC:* 0.8777; mean average error *MAE:* 0.2518; root-mean-square error *RMSE:* 0.3325 as a benchmark. In summary it can be said that all deep CNN models are superior to the shallow predictor with hand-crafted geometric feature or appearance feature (Liang et al., 2018).

Benchmark Dataset. The SCUT-FBP 5500 data set is a small data set for deep learning tasks. Therefore, it is an even greater challenge to train soft features like aesthetic or beauty. In order to measure the accuracy of the network and to be comparable to recent experiments in facial beauty prediction, we calculate the Pearson correlation coefficient (PC), mean absolute error (MAE) and root mean square error (RMSE).

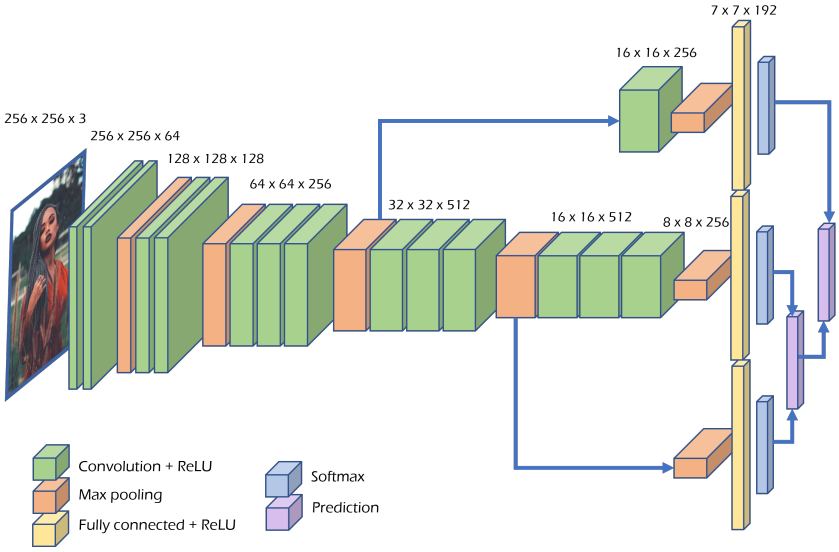

Figure 3: The architecture of AestheticNet is based on the VGG Face architecture and is expanded by two separate skip connections. At the end, the predictions of the differently convoluted feature vectors are added together.

AestheticNet predictor architecture. The VGG Face architecture (Simonyan & Zisserman, 2015) is the basis of our AestheticNet. Inspired by an idea of the paper from Shelhamer et al. (2017) we then add modifications to the network by exploiting feature maps from the third and fourth convolution block. Since the size of the features maps differ from the size of the resulting feature vector, we implement an additional max pooling layer to achieve the wanted output. For the predictions of the network, we concatenate the softmax results into a single feature vector as shown in fig. 3.

Our proposed network achieves a Pearson correlation coefficient of 0.9601, which indicates an almost linear correspondence between annotations and predictions. Our training results have a very high accuracy and outperform state-of-the-art results. The normalised mean square error is 3.896% and the normalised root mean square error is 5.580%. These are measurements of the average error of the predicted labels, which are used to evaluate the accuracy of the network. The results are normalised because there are different datasets with different score ranges.

Reannotation of SCUT-FBP5500 dataset. Since 2013, for our study of facial aesthetics, we conducted online surveys on multiple image datasets (mentioned in table 2) where thousands of students and their relatives participated. With this process we have been able to gather enough data to train a convolutional neural network with the goal to improve facial beauty prediction. During

Table 1: Comparison of prediction accuracy on SCUT-FBP5500

Architecture	PC	nMAE [%]	nRMSE [%]
AlexNet [1]	0.8298	7.345	9.548
AlexNet [2]	0.8634	n/a	n/a
ResNet-18 [3]	0.8513	7.045	9.258
ResNeXt-50 [4]	0.8777	6.295	8.313
HMTNet [5]	0.8783	6.2525	8.158
AaNet [6]	0.9055	5.590	7.385
P-AaNet [7]	0.8965	5.713	7.588
2M BeautyNet [8]	0.8996	n/a	n/a
EfficientNetB3 based AestheticNet (ours)	0.9011	5.841	7.663
VGG-Face based AestheticNet (ours)	0.9363	4.400	6.261
AestheticNet (ours)	**0.9601**	**3.896**	**5.580**

training convolutional neural networks (CNN) on this data, we recognised a large bias in this data. This led us to evaluate the annotations from Chinese and German universities and take a closer look at the bias. Our null hypothesis was that there is no bias in dependency of the ethical group, the proof for the presence of bias was done by reductio ad absurdum.

In null hypothesis significance testing, the p-value is the probability of obtaining test results at least as extreme as the results actually observed, under the assumption that the null hypothesis is correct (Aschwanden, 2015). The precise calculation of the p-value in this experiment is difficult because the factorials raise too high, to be reasonably computed on the thousands of labelled values. We calculate the p-value on 300 representative annotations which lead to a p-value of approximately 0.063%, therefore it is safe to say the null hypothesis can be rejected and we do have an ethical bias.

4 Training of Unbiased AI

In general, there are three main paths to reach the goal of unbiased predictions: fair pre-processing, fair in-processing and fair post-processing (Bellamy et al., 2018). Within this paper, we present two approaches based on those paths to train an unbiased network with biased data, for FBP. The first approach relies on data pre-processing before training to introduce fairness, we call it "balanced training". The second approach relies on a categorical cross entropy loss function, for the network to learn the bias and decrease it. Those processes are explained in the following sections.

4.1 Dataset and GAN images

Machine learning has evolved in the past decades and stands out due to the fact that the knowledge in the system is not provided by experts. Facial beauty prediction (FBP) that is consistent with human perception, is a significant visual recognition problem and a much-studied subject in recent decades. Eisenthal et al. (2006) and (Kagian et al., 2008) were among the first to publish their research about automatic facial attractiveness predictors and supervised learning techniques, based on the extraction of feature landmarks on faces. We analysed the data that we gathered with our Analysis Toolbox and could measure a significant bias within the prediction of aesthetics through different ethnicities. Therefore, training a network with the goal to create unbiased results is still a challenge in deep learning tasks. In the following we will first describe our data set blend and the accompanying Analysis Toolbox and we explain how we used a GAN to create artificial portraits with European and Asian ethnicities.

Starting in 2017, we used the Asian-European-dataset SCUT-FBP (Xie et al., 2015b; Liang et al., 2018) to evaluate biased annotations from Chinese and German universities. The results proved the assumption that German students favour images of European women and vice versa Chinese students rate Asian portraits higher. Since the SCUT-FBP 5500 dataset is a small dataset for deep learning tasks, we use data augmentation methods to enlarge the sample size of the training set by generating GAN images with either Asian or European or mixed images as input and new synthesised images as output. This augmentation method proves superior to geometric transformations like cropping

and rotating. All images are preprocessed, by normalisation methods to harmonise face pose, facial landmark positions and image size.

Table 2: Our dataset blend and annotations

datasets			annotations										
since	name	faces	age	gender	ethnic	height	weight	sports	glasses	attractiveness	complexion	hair colour	hair style
2013	MCSO Criminals [9]	750	✓	✓	✓	✓	✓	×	×	✓	×	✓	×
2013	Olympics [10]	1914	✓	✓	✓	✓	✓	✓	×	✓	×	×	×
2016	LFW [11]	1578	✓	×	✓	×	×	×	✓	✓	×	×	×
2018	SCUT-FBP* [12]	2750	✓	✓	✓	×	×	×	×	✓	×	×	×
2020	synthesised Eurasians (ours)	2942	✓	✓	✓	×	×	×	✓	✓	✓	✓	✓
2021	FairFace [13]	2100	✓	✓	✓	×	×	×	×	✓	×	×	×
		Σ 12034	✓		✓					✓			

For the purpose of a thorough analysis, we blend multiple datasets in the domain of facial aesthetics together. Our complete set of databases which is described in table 2, consists of multiracial and multiethnic individuals. In total, this data set includes 12,034 portrait images from persons of different ethnicities with individual social backgrounds. These images are labelled and annotated in surveys over a period of 8 years with a total number of 5.4 million annotations. Additionally, recently we add the FairFace (Kärkkäinen & Joo, 2019) database, which includes male and female portraits of seven different ethnic groups.

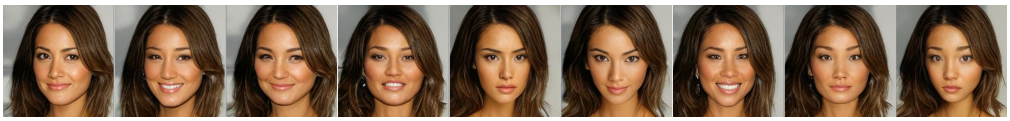

Figure 4: StarGAN v2 generated Eurasians. From left to right: 90%, 80%, 70%, 60% European, half/half, 60%, 70%, 80% 90% Asian

The synthesised Eurasians images are artificially generated with StarGAN v2 (Choi et al., 2020) to determine the influence of the biased view of annotators on aesthetics of persons from different ethnicities. We used different customised input for the source and reference images to control the amount of ethnic admixture. Figure 4 shows one exemplary set of images for the Eurasians dataset.

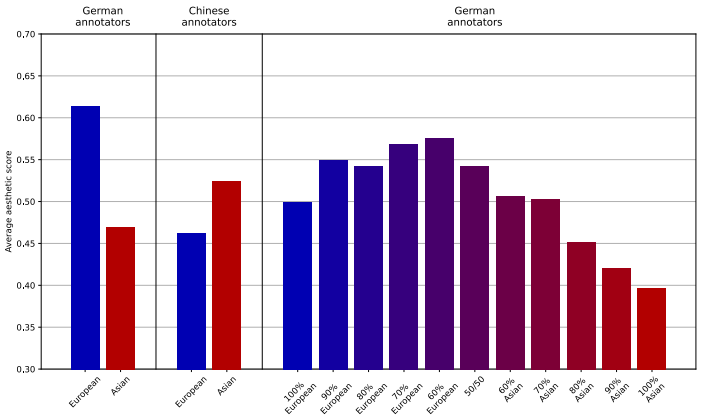

Figure 5: Unconscious bias towards ethnic aesthetic of either German or Chinese annotators. Left: average aesthetic score on SCUT-FBP by German annotators, middle: average aesthetic score labelled by Chinese students, right: aesthetic scores on the Eurasian dataset annotated by German students.

After annotating the dataset, the unconscious bias in the annotations can be uncovered. Figure 5 shows the biased average score of our networks on the SCUT-FBP dataset and the Eurasian dataset.

Figure 6 illustrates the analysis on the distribution of aesthetic score and age for Asians, Europeans and three mixed-racial subgroups. The different group annotation points are displayed in different colours. We calculate the following metrics for each group cluster i: Horizontal dashed lines are average attractiveness values $\overline{a}_i$. Vertical dashed lines are average age values $\overline{y}_i$. As can be seen, the interval of $\overline{a}_i$ has a small span, yet however the interval of $\overline{y}_i$ has a significantly larger span. Each $\overline{a}_i$ and $\overline{y}_i$ values intersection point forms an per group attractiveness-age-factor $AAF_i = \overline{a}_i/\overline{y}_i$. In a fair machine, these AAF_i points would be closer together, as the $\overline{y}_i$ span is small. This idea is further elaborated in section 4.

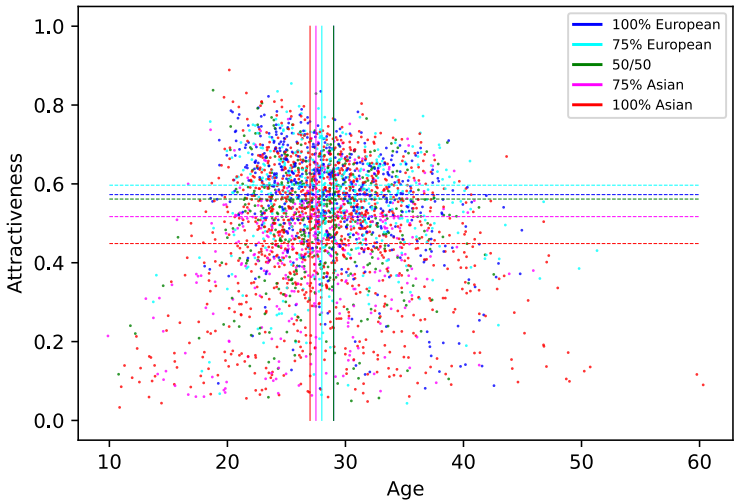

Figure 6: Biased correlation between attractiveness, age and ethnicity by German annotators. In an ethical, fair network the attractiveness for equal age groups would be the same. This would be represented in the figure by the same height of the lines for equal age groups.

4.2 Training and Data Pre-Processing

In our first approach of training the network we have applied pre-processing and resampling to the input data, which is explained in the following paragraphs.

This paper proposes a way to create a fair network with this biased data. Therefore, the bias must be identified in the ground truth labels of the dataset and divided into two subsets. The first subset (German annotations) confirms and increases the existing bias whereas the second subset (Chinese annotations) consists of the contrary prejudices. Afterwards, a GAN then generates synthetic images, which are a gradation of the mixture of the first and second subset. The least biased result according to our understanding is the best balance of the generated images. This knowledge can then be applied back to the original data set. This implies the height difference of all the bars should be minimised.

In our training process, we have a clear bias in the annotations, as shown in fig. 2 and measured in the analysis of the data. If we train our network based only on this labels, it follows the data and replicates the bias from the annotations, as shown in the comparison of the predictions with the annotations in fig. 2. Chinese annotators rate Asian faces higher, based on this data our prediction is biased towards higher aesthetic scores of Asian Faces. This is the same if we train the machine only on European Faces, annotated by Germans. In the next training, we added the annotations from the Chinese and German annotators and trained the network on an equal distribution of those annotations (Ratio: 1.0). The result is shown in fig. 7 on the left side of the diagram. The average aesthetic rating of European and Asian faces is still biased, however not as strong as in the previous experiment. The eleven bars on the right side of fig. 7 show the average aesthetic score based on the ethnicity. The bias is shown in more detail, ranging from 100% to 70% European who have the highest aesthetic score, to the lowest aesthetic score, the more Asian looking the portrait is. As a result, the network is still biased, a network trained on this data reflects the bias in the FBP.

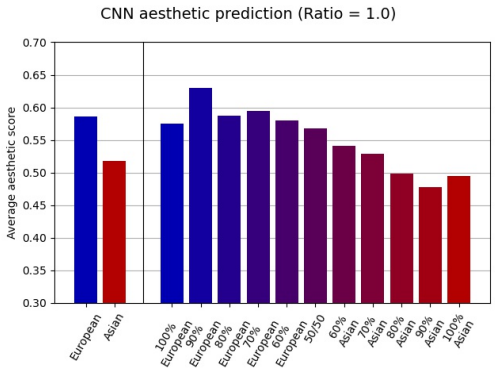

Figure 7: FBP with same amount of Asian and German labels. Ratio = 1.0 stands for the same weight ω for German and Asian annotations.

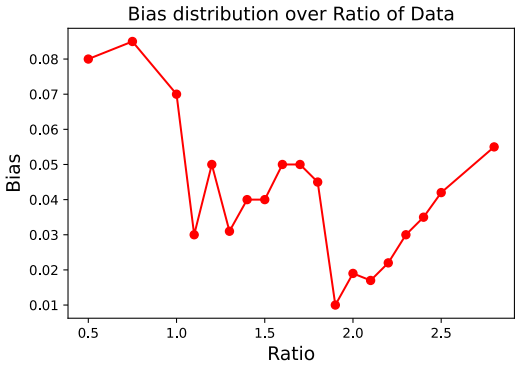

Figure 8: Correlation of the bias over the ratio of German and Chinese annotations. The least bias here is at the ratio of 1.9

In this experiment, balancing the training data means to find the minimum by concatenating the German annotated subset g with the weighted ω Chinese annotated subset c. The goal in this approach is to level the average aesthetic scores $\overline{g}$ and $\overline{c}$ for the generated predictions g_i and c_i. The network bias B is then defined by

$$B = \frac{1}{2n+1} \sum_{i=0}^{n} |\overline{g} - g_i| + \omega \, |\overline{c} - c_i|. \tag{1}$$

Starting from a ratio of 1:1, in which German and Chinese annotations are distributed equally, we gradually increase the weight of the Chinese annotations. Technically, the balancing of distribution of the training data is done with a factor based approach. First, the ratio between Chinese and European annotations are calculated. Secondly, the factor for the balanced distribution is determined in a stochastic approach. In our experiment we varied the ratio from 2:1 to 1:3.2 for German annotations to Chinese annotations. Each training step and the corresponding bias over the ratio is shown in fig. 8. Determining the minimum in fig. 8 is equal to finding the least biased network. It is visible that a ratio of 1:1.9 produces the least biased network for this experiment and its results are shown in fig. 9. This means the Chinese annotations are weighted nearly double the amount than the European annotations.

Limitations of this approach are that information about the structure of the underlying latent features are unknown and balancing the network requires a lot of time and work. Therefore, we additionally propose another approach, described in the following section.

4.3 Debiasing Neural Network

4.3.1 Training Network Features

Regular convolutional neural networks (CNN) are generally used for face recognition tasks and we also used CNNs for FBP. They can be used to classify identities, and in our case to classify aesthetic scores (Serengil & Ozpinar, 2020), commonly called Facial Beauty Prediction (FBP). FBP using Machine Learning and Artificial Intelligence has been researched and improved many times in the past by various researches (Eisenthal et al., 2006; Gerlach et al., 2020; Kagian et al., 2008; Xie et al., 2015b; Liang et al., 2018; Liu et al., 2016; Xu et al., 2019).

Common for all those studies is that data is often generated by subgroups, with their own characteristics and behaviours (Mehrabi et al., 2019), especially in the highly subjective field of aesthetic rating. Therefore, the possibility exists, that all datasets are affected by bias, which the networks trained on them transfer into the FBP. Solving this problem, training an unbiased network with biased data, is a recently much discussed subject in the area of Machine Learning (Amini et al., 2019; Bellamy et al., 2018).

To achieve the first results on unbiased aesthetic estimation, we used the existing VGG-Face framework in Keras with TensorFlow and adjusted it. The network consists of 11 blocks, each containing a

linear operator and followed by one or more non-linearities such as ReLU and max pooling (Parkhi et al., 2015). We apply transfer learning here and use the pretrained model for Face Recognition (Parkhi et al., 2015). Building up on the face recognition, attractiveness estimation is similar to age estimation (Gyawali et al., 2020) performed by observing the facial features from portraits. Comparable to age estimation, the network then assigns the Portrait a beauty score.

The convolutional layers in the network are followed by a rectification layer (ReLu) as in (Krizhevsky et al., 2017). We used the Adam optimizer (Kingma & Ba, 2017). The input to our network is a face image of the size $256 \times 256 \times 3$, and it uses Zero-Padding around the edges, to ensure that the image information on the edge is not lost. Our input data is split into 60% train and 40% test data. The convolutional layers parameters of VGG-Face are not changed and kept frozen during the training. We use a dropout of 50%, and as it is a regression problem our final layer must be the size of 1. To classify the aesthetic score, a softmax activation function is used in the final layer. As a loss metric, we use the mean squared error and to compare our networks we also calculate the Pearson correlation and the root mean squared error.

4.3.2 Balanced Training

The process and the effect of the ratio on the average aesthetic rating is shown in fig. 8. By modifying the ratio of the annotations a minimum is determined that illustrates the lowest difference between the average aesthetic prediction of Asian and European faces. This represents a specific loss function for our network that maps bias onto measurable values. To remove bias from our network, we calculate the difference between European and Asian aesthetic predictions and find the global minimum.

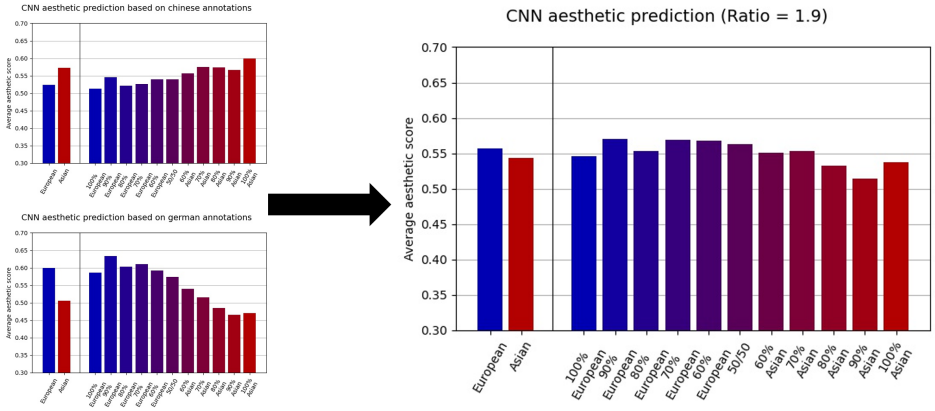

Figure 9: CNN aesthetic prediction with equalised distribution of training data. The charts on the left side show the prediction of the network if it is only trained on Chinese or German annotations. On the right side, the prediction of the network, which was trained on the biased data is shown. All bars have more or less the same height and only differ minimally. This means, that we could eliminate most of the bias in the training data, by balancing and we can assume that this trained network is fair.

The minimum of the average aesthetic score between Asian and European faces is located at a ratio of 1:1.9 where the average aesthetic score differs by about 5%. We create a model with a fair performance over all classes of different ethnicities as shown in fig. 9. This proves, that by resampling and balancing the training data a less biased AI can be created. We are retaining the precision in FBP, as shown in table 1. This process creates a less-biased AI in FBP tasks.

Our results are displayed in fig. 9 where all bar charts have a similar height and the FBP score is considerably less biased. Not all bars have the exact same height, this is due to some background noise. Real world data usually contains noise which affects tasks such as classification in machine learning (Gupta & Gupta, 2019). This noise also affects our aesthetic prediction, however with those minor differences, we can consider our network as unbiased and therefore fair. As we use a factor based approach to multiply the annotation data, this noise is present over all ratios. Only the difference of the averages increases or decreases within the variations of the ratio.

5 CONCLUSION

Our two main contributions are AestheticNet and a new approach to bias-free machine learning tools. In this work, we have proposed to augment the SCUT-FBP dataset by synthesised GAN images and show that AestheticNet predicts facial attractiveness with higher correlation then competitive approaches. Then we utilise a novel learning strategy to minimise bias in networks. Unbiased networks are an important step towards a future, where more decisions are made by AI and therefore more lives are influenced by artificial intelligence - unbiased decision making is the foundation of ethical and moral values.

Bias-free decision making is a challenging problem in machine learning tasks, yet it yields the great potential to be one of the most significant strengths of an AI. We have shown a method to eliminate bias in facial attractiveness prediction and this method can be transferred to multiple similar networks.

Training an unbiased model on biased data is an important goal from a Machine Learning Perspective, as perfect, unbalanced data might be raw. Especially in the field of Aesthetic Judgement, it is important that the machine is able to realise the bias. By learning, how to act against this bias, we can scale this approach in the future on larger datasets in other areas. The algorithm is introduced by applying it to aesthetic judgement, but not limited to it. Further development and deployment of fair and unbiased AI systems is crucial for AI to be a social benefit for all and reduce algorithmic discrimination.

Implicit bias has always been a hot-spot in the field of psychology in the 21st century. With the intersection of disciplines, a series of moral and ethical issues arising from it have also attracted the attention of the philosophical field. Implicit bias is widespread and is in a silent way. It affects all aspects of our lives, even in the field of artificial intelligence, which is equally popular in the 21st century. In this article, we have verified the universality of implicit existence and artificial intelligence will copy human prejudice by proving the establishment of two hypotheses. On this basis, we have analysed the reasons for the existence of implicit bias from neuroscience and psychology. Next, we used experimental methods to systematically demonstrate how human implicit bias affects the decision-making of artificial intelligence and found a way to eliminate the implicit bias of artificial intelligence. Secondly, we improved the fairness of the algorithm from machine learning. From a perspective, the training of unbiased models on biased data is an important goal, and unbiased networks are an important step towards the future. In the future, artificial intelligence will make more decisions, so more lives will be affected by artificial intelligence. Unbiased artificial intelligence decision-making is the moral foundation.

One problem that has to be tackled in the future is the issue that our attractiveness was elo rated and is hence not equally distributed per level of attractiveness. Very attractive and very unattractive faces are much less common than average faces. In order to make up for that we used stargan_v2 (Choi et al., 2020) to enhance our data set by computer generated faced that are meant to be more attractive than real faces. The generated images were rated in another survey and in fact came out to be more attractive than real ones. As scientific AI-researchers, we require our work to maintain sufficient awe of nature and morality. This is our current and future work. As Kant said, "Two things fill the mind with ever new and increasing admiration and awe, the more often and steadily we reflect upon them: the starry heavens above me and the moral law within me."

REFERENCES

Frances E. Aboud. *Children and prejudice*. Social psychology and society. B. Blackwell, Oxford, OX, UK ; Cambridge, MA, USA, 1989. ISBN 978-0-631-14939-2 978-0-631-14941-5.

Ali Akbari, Muhammad Awais, Zhen-Hua Feng, Ammarah Farooq, and Josef Kittler. A flatter loss for bias mitigation in cross-dataset facial age estimation. *CoRR*, abs/2010.10368, 2020. URL https://arxiv.org/abs/2010.10368.

Alexander Amini, Ava P. Soleimany, Wilko Schwarting, Sangeeta N. Bhatia, and Daniela Rus. Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 289–295, Honolulu HI USA, January 2019. ACM. ISBN 978-1-4503-6324-2. doi: 10.1145/3306618.3314243. URL https://dl.acm.org/doi/10.1145/3306618.3314243.

Petra Anýžová and Petr Matějů. Beauty still matters: The role of attractiveness in labour market outcomes. *International Sociology*, 33(3):269–291, May 2018. ISSN 0268-5809, 1461-7242. doi: 10.1177/0268580918760431. URL http://journals.sagepub.com/doi/10.1177/0268580918760431.

Christie Aschwanden. Not even scientists can easily explain p-values. *FiveThirtyEight. com, Nov*, 24: 2015, 2015.

Simon Baron-Cohen. *The essential difference*. Penguin, London, 2012. ISBN 978-0-241-96135-3. OCLC: 809151983.

BBC. Google apologises for Photos app's racist blunder. *BBC News*, July 2015. URL https://www.bbc.com/news/technology-33347866.

Geoffrey Beattie and Patrick Johnson. Possible unconscious bias in recruitment and promotion and the need to promote equality. *Perspectives: Policy and Practice in Higher Education*, 16(1): 7–13, January 2012. ISSN 1360-3108, 1460-7018. doi: 10.1080/13603108.2011.611833. URL http://www.tandfonline.com/doi/abs/10.1080/13603108.2011.611833.

beauty.ai. The First International Beauty Contest Judged by Artificial Intelligence, 2016. URL http://beauty.ai.

Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. *arXiv:1810.01943 [cs]*, October 2018. URL http://arxiv.org/abs/1810.01943. arXiv: 1810.01943.

Miranda Bogen. All the Ways Hiring Algorithms Can Introduce Bias. *Harvard Business Review*, May 2019. ISSN 0017-8012. URL https://hbr.org/2019/05/all-the-ways-hiring-algorithms-can-introduce-bias. Section: Hiring.

Nick Bostrom and Eliezer Yudkowsky. The Ethics of Artificial Intelligence. *Chapman and Hall*, pp. 21, 2018.

Jessica A. Cameron, Jeannette M. Alvarez, Diane N. Ruble, and Andrew J. Fuligni. Children's Lay Theories About Ingroups and Outgroups: Reconceptualizing Research on Prejudice. *Personality and Social Psychology Review*, 5(2):118–128, May 2001. ISSN 1088-8683. doi: 10.1207/S15327957PSPR0502_3. URL https://doi.org/10.1207/S15327957PSPR0502_3. Publisher: SAGE Publications Inc.

Fernanda Carrera. Race and gender of aesthetics and affections: algorithmization of racism and sexism in contemporary digital image databases. *Matrizes*, 14(2):217–240, 2020.

John Cawley. The Impact of Obesity on Wages. *Journal of Human Resources*, XXXIX(2):451–474, 2004. ISSN 0022-166X, 1548-8004. doi: 10.3368/jhr.XXXIX.2.451. URL http://jhr.uwpress.org/lookup/doi/10.3368/jhr.XXXIX.2.451.

Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. StarGAN v2: Diverse Image Synthesis for Multiple Domains. *arXiv:1912.01865 [cs]*, April 2020. URL http://arxiv.org/abs/1912.01865. arXiv: 1912.01865.

Xuan Cui, Qiuping Cheng, Wuji Lin, Jiabao Lin, and Lei Mo. Different influences of facial attractiveness on judgments of moral beauty and moral goodness. *Scientific Reports*, 9 (1), December 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-48649-5. URL http://www.nature.com/articles/s41598-019-48649-5.

Jeffrey Dastin. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*, October 2018. URL https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.

Hume David. *Of the Standard of Taste*. Essays Moral, Political and Literary. Longmans, Green and Co., London, 1898. edited by T.H. Green and T. H. Grose.

J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, June 2009. doi: 10.1109/CVPR.2009.5206848. ISSN: 1063-6919.

Eric Dietrich. Homo Sapiens 2.0 Why We Should Build the Better Robots of Our Nature. In M. Anderson S. Anderson (ed.), *Machine Ethics*. Cambridge Univ. Press, 2011.

F. Dornaika, A. Moujahid, K. Wang, and X. Feng. Efficient deep discriminant embedding: Application to face beauty prediction and classification. *Engineering Applications of Artificial Intelligence*, 95:103831, 2020. ISSN 0952-1976. doi: https://doi.org/10.1016/j.engappai.2020.103831. URL https://www.sciencedirect.com/science/article/pii/S0952197620302013.

Yael Eisenthal, Gideon Dror, and Eytan Ruppin. Facial Attractiveness: Beauty and the Machine. *Neural Computation*, 18(1):119–142, January 2006. ISSN 0899-7667, 1530-888X. doi: 10.1162/089976606774841602. URL https://www.mitpressjournals.org/doi/abs/10.1162/089976606774841602.

European Commission High-Level Expert Group on Artificial Intelligence [AI HLEG]. A definition of AI. Technical report, European Union, Robotics and Artificial Intelligence Innovation and Excellence (Unit A.1), 2019. URL https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

Facebook Inc. Managing Bias, 2014. URL https://managingbias.fb.com/.

Michael T. French. Physical appearance and earnings: further evidence. *Applied Economics*, 34(5):569–572, March 2002. ISSN 0003-6846, 1466-4283. doi: 10.1080/00036840010027568. URL http://www.tandfonline.com/doi/abs/10.1080/00036840010027568.

Junying Gan, Fabio Scotti, Li Xiang, Yikui Zhai, Chaoyun Mai, Guohui He, Junying Zeng, Zhenfeng Bai, Ruggero Donida Labati, and Vincenzo Piuri. 2m beautynet: Facial beauty prediction based on multi-task transfer learning. *IEEE Access*, 8:20245–20256, 2020. doi: 10.1109/ACCESS.2020.2968837. URL https://doi.org/10.1109/ACCESS.2020.2968837.

Tobias Gerlach, Michael Danner, Le Peng, Aidas Kaminickas, Wu Fei, and Matthias Rätsch. Who Loves Virtue as much as He Loves Beauty?: Deep Learning based Estimator for Aesthetics of Portraits:. In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pp. 521–528, Valletta, Malta, 2020. SCITEPRESS - Science and Technology Publications. ISBN 978-989-758-402-2. doi: 10.5220/0009172905210528. URL http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0009172905210528.

Google LLC. re:Work - Guide: Raise awareness about unconscious bias, 2016. URL https://rework.withgoogle.com/guides/unbiasing-raise-awareness/steps/watch-unconscious-bias-at-work/.

Richard D. Gross. *Psychology: the science of mind and behaviour*. Hodder Education, London, 6th ed edition, 2010. ISBN 978-1-4441-0831-6.

Shivani Gupta and Atul Gupta. Dealing with Noise Problem in Machine Learning Data-sets: A Systematic Review. *Procedia Computer Science*, 161:466–474, January 2019. ISSN 1877-0509. doi: 10.1016/j.procs.2019.11.146. URL https://www.sciencedirect.com/science/article/pii/S1877050919318575.

D. Gyawali, P. Pokharel, A. Chauhan, and S. C. Shakya. Age Range Estimation Using MTCNN and VGG-Face Model. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pp. 1–6, July 2020. doi: 10.1109/ICCCNT49239.2020.9225443.

Jonathan Haidt. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108(4):814–834, 2001. ISSN 1939-1471, 0033-295X. doi: 10.1037/0033-295X.108.4.814. URL http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-295X.108.4.814.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385.

Evelyn Hello, Peer Scheepers, and Mérove Gijsberts. Education and Ethnic Prejudice in Europe: Explanations for cross-national variances in the educational effect on ethnic prejudice. *Scandinavian Journal of Educational Research*, 46(1):5–24, March 2002. ISSN 0031-3831, 1470-1170. doi: 10.1080/00313830120115589. URL http://www.tandfonline.com/doi/abs/10.1080/00313830120115589.

Evan Hill, Ainara Tiefenthäler, Christiaan Triebert, Drew Jordan, Haley Willis, and Robin Stein. How George Floyd Was Killed in Police Custody. *The New York Times*, June 2020. ISSN 0362-4331. URL https://www.nytimes.com/2020/05/31/us/george-floyd-investigation.html.

Eric Horvitz. AI, people, and society. *Science*, 357(6346):7–7, July 2017. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aao2466. URL https://www.sciencemag.org/lookup/doi/10.1126/science.aao2466.

Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

R. Hülsmann. MCSO Online Inmate Data, 2013. URL https://www.mcso.us/PAID/Home/SearchResults.

R. Hülsmann and S. Braun. *Olympic Photos, Galleries and Slideshows*. https://www.olympic.org/photos, Retrieved 03/08/2013.

Jill Billante, Chuck Haddad, and Margaret Beale Spencer. Study: White and black children biased toward lighter skin - CNN.com, 2010. URL http://www.cnn.com/2010/US/05/13/doll.study/index.html.

Michael Johns, Toni Schmader, and Andy Martens. Knowing Is Half the Battle: Teaching Stereotype Threat as a Means of Improving Women's Math Performance. *Psychological Science*, 16(3):175–179, March 2005. ISSN 0956-7976, 1467-9280. doi: 10.1111/j.0956-7976.2005.00799.x. URL http://journals.sagepub.com/doi/10.1111/j.0956-7976.2005.00799.x.

Amit Kagian, Gideon Dror, Tommer Leyvand, Isaac Meilijson, Daniel Cohen-Or, and Eytan Ruppin. A machine learning predictor of facial attractiveness revealing human-like psychophysical biases. *Vision Research*, 48(2):235–243, January 2008. ISSN 00426989. doi: 10.1016/j.visres.2007.11.007. URL https://linkinghub.elsevier.com/retrieve/pii/S0042698907005032.

Daniel Kahneman. *Thinking, fast and slow*. Farrar, Straus and Giroux, New York, 1st ed edition, 2011. ISBN 978-0-374-27563-1 978-0-374-53355-7 978-0-606-27564-4.

Immanuel Kant and Paul Guyer. *Critique of the Power of Judgment*. Cambridge University Press, 1 edition, September 2000. ISBN 978-0-521-34892-8 978-0-521-34447-0 978-0-511-80465-6. doi: 10.1017/CBO9780511804656. URL https://www.cambridge.org/core/product/identifier/9780511804656/type/book.

Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. *arXiv:1812.04948 [cs, stat]*, March 2019. URL http://arxiv.org/abs/1812.04948. arXiv: 1812.04948.

Jozef Kelemen, Jan Romportl, and Eva Zackova (eds.). *Beyond Artificial Intelligence: The Disappearing Human-Machine Divide*. Number 9 in Topics in Intelligent Engineering and Informatics. Springer International Publishing : Imprint: Springer, Cham, 1st ed. 2015 edition, 2015. ISBN 978-3-319-09668-1. doi: 10.1007/978-3-319-09668-1.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*, January 2017. URL http://arxiv.org/abs/1412.6980. arXiv: 1412.6980.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, May 2017. ISSN 0001-0782, 1557-7317. doi: 10.1145/3065386. URL https://dl.acm.org/doi/10.1145/3065386.

Kimmo Kärkkäinen and Jungseock Joo. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age. *arXiv:1908.04913 [cs]*, August 2019. URL http://arxiv.org/abs/1908.04913. arXiv: 1908.04913.

Calvin K. Lai and Edmond J. Reducing implicit racial preferences: II. Intervention effectiveness across time., 2016.

Francisco Lara and Jan Deckers. Artificial Intelligence as a Socratic Assistant for Moral Enhancement. *Neuroethics*, 13(3):275–287, October 2020. ISSN 1874-5504. doi: 10.1007/s12152-019-09401-y. URL https://doi.org/10.1007/s12152-019-09401-y.

Jeff Larson and Julia Angwin. Bias in Criminal Risk Scores Is Mathematically Inevitable, Researchers Say, 2016. URL www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say.

Sam Levin. A beauty contest was judged by AI and the robots didn't like dark skin, September 2016. URL http://www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people. Section: Technology.

Lingyu Liang, Luojun Lin, Lianwen Jin, Duorui Xie, and Mengru Li. SCUT-FBP5500: A Diverse Benchmark Dataset for Multi-Paradigm Facial Beauty Prediction. *arXiv:1801.06345 [cs]*, January 2018. URL http://arxiv.org/abs/1801.06345. arXiv: 1801.06345.

Luojun Lin, Lingyu Liang, Lianwen Jin, and Weijie Chen. Attribute-aware convolutional neural networks for facial beauty prediction. In Sarit Kraus (ed.), *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pp. 847–853. ijcai.org, 2019. doi: 10.24963/ijcai.2019/119. URL https://doi.org/10.24963/ijcai.2019/119.

Shu Liu, Bo Li, Yangyu Fan, Zhe Guo, and Ashok Samal. Label distribution based facial attractiveness computation by deep residual learning. *arXiv:1609.00496 [cs]*, September 2016. URL http://arxiv.org/abs/1609.00496. arXiv: 1609.00496.

Kirsten Lloyd. Bias amplification in artificial intelligence systems, 2018.

Keith B. Maddox and Jennifer M. Perry. Racial Appearance Bias: Improving Evidence-Based Policies to Address Racial Disparities. *Policy Insights from the Behavioral and Brain Sciences*, 5 (1):57–65, March 2018. ISSN 2372-7322, 2372-7330. doi: 10.1177/2372732217747086. URL http://journals.sagepub.com/doi/10.1177/2372732217747086.

Andrew Mason. Appearance, Discrimination, and Reaction Qualifications. *Journal of Political Philosophy*, 25(1):48–71, 2017. ISSN 1467-9760. doi: https://doi.org/10.1111/jopp.12099. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/jopp.12099. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jopp.12099.

Kabir Matharu, JohannaF Shapiro, RachelR Hammer, Rl Kravitz, MachelleD Wilson, and FaithT Fitzgerald. Reducing obesity prejudice in medical education. *Education for Health*, 27 (3):231, 2014. ISSN 1357-6283. doi: 10.4103/1357-6283.152176. URL http://www.educationforhealth.net/text.asp?2014/27/3/231/152176.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A Survey on Bias and Fairness in Machine Learning. *arXiv:1908.09635 [cs]*, September 2019. URL http://arxiv.org/abs/1908.09635. arXiv: 1908.09635.

C. A. Moss-Racusin, J. F. Dovidio, V. L. Brescoll, M. J. Graham, and J. Handelsman. Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, 109 (41):16474–16479, October 2012. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1211286109. URL http://www.pnas.org/cgi/doi/10.1073/pnas.1211286109.

Jean Moule. Understanding Unconscious Bias and Unintentional Racism. *Phi Delta Kappan*, 90(5):320–326, January 2009. ISSN 0031-7217. doi: 10.1177/003172170909000504. URL https://doi.org/10.1177/003172170909000504. Publisher: SAGE Publications Inc.

Siddhartha Mukherjee. *Gesetze der Medizin: Anmerkungen zu einer ungewissen Wissenschaft*. TEDBooks. Fischer Taschenbuch, Frankfurt am Main, 2016. ISBN 978-3-596-03468-0. OCLC: 949778371.

Kevin Munger. Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment. *Political Behavior*, 39(3):629–649, September 2017. ISSN 0190-9320, 1573-6687. doi: 10.1007/s11109-016-9373-5. URL http://link.springer.com/10.1007/s11109-016-9373-5.

A. B. C. News. What a Doll Tells Us About Race, 2009. URL https://abcnews.go.com/GMA/story?id=7213714&page=1.

Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep Face Recognition. In *Procedings of the British Machine Vision Conference 2015*, pp. 41.1–41.12, Swansea, 2015. British Machine Vision Association. ISBN 978-1-901725-53-7. doi: 10.5244/C.29.41. URL http://www.bmva.org/bmvc/2015/papers/paper041/index.html.

Derek S. Prijatelj, Mel McCurrie, and Walter J. Scheirer. A Bayesian Evaluation Framework for Ground Truth-Free Visual Recognition Tasks. *arXiv:2007.06711 [cs, stat]*, June 2020. URL http://arxiv.org/abs/2007.06711. arXiv: 2007.06711.

Stephen H. Richmond. The Beholder's I: The Perception of Beauty and the Development of the Self. *Perception of Beauty*, October 2017. doi: 10.5772/intechopen.69531. URL https://www.intechopen.com/books/perception-of-beauty/the-beholder-s-i-the-perception-of-beauty-and-the-development-of-the-self. Publisher: IntechOpen.

Dustin Rynders. Battling Implicit Bias in the IDEA to Advocate for African American Students with Disabilities. *Touro Law Review*, 35(1), January 2019. ISSN 8756-7326. URL https://digitalcommons.tourolaw.edu/lawreview/vol35/iss1/18.

Joel Schwarz. Roots of unconscious prejudice affect 90 to 95 percent of people, psychologists demonstrate at press conference — UW News. https://www.washington.edu/news/1998/09/29/, September 1998. (Accessed on 04/24/2021).

S. I. Serengil and A. Ozpinar. LightFace: A Hybrid Deep Face Recognition Framework. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pp. 1–5, October 2020. doi: 10.1109/ASYU50717.2020.9259802.

Daniel B. Shank and Alyssa DeSanti. Attributions of morality and mind to artificial intelligence after real-world moral violations. *Computers in Human Behavior*, 86:401–411, 2018. ISSN 0747-5632. doi: https://doi.org/10.1016/j.chb.2018.05.014. URL https://www.sciencedirect.com/science/article/pii/S0747563218302401.

Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D. Sculley. No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World. *arXiv:1711.08536 [stat]*, November 2017. URL http://arxiv.org/abs/1711.08536. arXiv: 1711.08536.

Shubham Sharma, Yunfeng Zhang, Jesús M. Ríos Aliaga, Djallel Bouneffouf, Vinod Muthusamy, and Kush R. Varshney. Data Augmentation for Discrimination Prevention and Bias Disambiguation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, pp. 358–364, New York, NY, USA, February 2020. Association for Computing Machinery. ISBN 978-1-4503-7110-0. doi: 10.1145/3375627.3375865. URL https://doi.org/10.1145/3375627.3375865.

Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):640–651, 2017. doi: 10.1109/TPAMI.2016.2572683. URL https://doi.org/10.1109/TPAMI.2016.2572683.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1409.1556.

Stacey Sinclair, Elizabeth Dunn, and Brian Lowery. The relationship between parental racial attitudes and children's implicit prejudice. *Journal of Experimental Social Psychology*, 41(3):283–289, May 2005. ISSN 00221031. doi: 10.1016/j.jesp.2004.06.003. URL https://linkinghub.elsevier.com/retrieve/pii/S0022103104000666.

Seth Stephens-Davidowitz. The cost of racial animus on a black candidate: Evidence using Google search data. *Journal of Public Economics*, 118:26–40, October 2014. ISSN 00472727. doi: 10.1016/j.jpubeco.2014.04.010. URL https://linkinghub.elsevier.com/retrieve/pii/S0047272714000929.

Björn Wallace, David Cesarini, Paul Lichtenstein, and Magnus Johannesson. Heritability of ultimatum game responder behavior. *Proceedings of the National Academy of Sciences*, 104(40):15631–15634, October 2007. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0706642104. URL https://www.pnas.org/content/104/40/15631. ISBN: 9780706642100 Publisher: National Academy of Sciences Section: Social Sciences.

Daniël H. J. Wigboldus, Jeffrey W. Sherman, Heather L. Franzese, and Ad van Knippenberg. Capacity and Comprehension: Spontaneous Stereotyping Under Cognitive Load. *Social Cognition*, 22 (3):292–309, June 2004. ISSN 0278-016X. doi: 10.1521/soco.22.3.292.35967. URL http://guilfordjournals.com/doi/10.1521/soco.22.3.292.35967.

Duorui Xie, Lingyu Liang, Lianwen Jin, Jie Xu, and Mengru Li. SCUT-FBP: A benchmark dataset for facial beauty perception. In *2015 IEEE International Conference on Systems, Man, and Cybernetics, Kowloon Tong, Hong Kong, October 9-12, 2015*, pp. 1821–1826. IEEE, 2015a. doi: 10.1109/SMC.2015.319. URL https://doi.org/10.1109/SMC.2015.319.

Duorui Xie, Lingyu Liang, Lianwen Jin, Jie Xu, and Mengru Li. SCUT-FBP: A Benchmark Dataset for Facial Beauty Perception. *arXiv:1511.02459 [cs]*, November 2015b. URL http://arxiv.org/abs/1511.02459. arXiv: 1511.02459.

L. Xu, H. Fan, and J. Xiang. Hierarchical Multi-Task Network For Race, Gender and Facial Attractiveness Recognition. In *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 3861–3865, September 2019. doi: 10.1109/ICIP.2019.8803614. ISSN: 2381-8549.

Matthew D. Zeiler and Rob Fergus. Visualizing and Understanding Convolutional Networks. *arXiv:1311.2901 [cs]*, November 2013. URL http://arxiv.org/abs/1311.2901. arXiv: 1311.2901.

Yikui Zhai, He Cao, Wenbo Deng, Junying Gan, Vincenzo Piuri, and Jun-Ying Zeng. Beautynet: Joint multiscale CNN and transfer learning method for unconstrained facial beauty prediction. *Comput. Intell. Neurosci.*, 2019:1910624:1–1910624:14, 2019. doi: 10.1155/2019/1910624. URL https://doi.org/10.1155/2019/1910624.

M. Zhang and G. Kreiman. Beauty is in the eye of the machine. In *Nat Hum Behav 5, 675–676 ()*., 2021. doi: https://doi.org/10.1038/s41562-021-01125-5.

Frederik Zuiderveen Borgesius. Discrimination, artificial intelligence, and algorithmic decision-making. *Directorate General of Democracy © Council of Europe*, pp. 51, 2018. URL https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73.%20Page%2010.

Piotr Żuk. Nation, national remembrance, and education — Polish schools as factories of nationalism and prejudice. *Nationalities Papers*, 46(6):1046–1062, November 2018. ISSN 0090-5992, 1465-3923. doi: 10.1080/00905992.2017.1381079. URL https://www.cambridge.org/core/product/identifier/S0090599200043828/type/journal_article.

A SUPPLEMENTARY MATERIAL

Toolbox The proposed toolbox introduces a huge collection of machine learning and face analysis applications. Specifically, to evaluate the annotation data for discriminatory bias, we designed the application that contains dedicated modules to detect correlations inside the network. The toolbox itself consists of the following modules: (a) preparation of the annotated data, (b) statistics generator, (c) dataset administration, (d) pretrained convolutional neural network (CNN) for hair colour, hair style, skin complexion, (e) unbiased, non-discriminatory aesthetic scores, (f) facial landmark detector, (g) 3D morphable model fitter.

Pearson Correlation Coefficient The Pearson correlation coefficient is a value between -1 and 1, which is used to measure the correlation (linear correlation) between two variables X and Y; when one variable increases, the other variable also increases, indicating that there is a positive correlation between them, and the correlation coefficient is greater then 0; if one variable increases, the other variable decreases, indicating that there is a negative correlation between them, and the correlation coefficient is less than 0; if the correlation coefficient is equal to 0, there is no linearity correlation relationship, if the correlation coefficient is equal to 1, it means that they are linearly equal, that is, the scoring results after machine learning are completely equivalent to the artificial results of the experimental participants in China and Germany.

The Pearson Correlation Coefficient (PC) is a statistic that measures linear correlation between the annotation and the FBP of AestheticNet. The value of the PC is in the range of 1 to -1. 1 or -1 means there is a high linear correlation. 0 means there is no correlation.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2}$$

nMAE To compare different approaches we needed to normalise the errors. With this normalisation we are able to compare between datasets or models with different score ranges (e.g. beauty score from 1-5 or 1-7) with ours, as we used a score range from 1-10. The error is expressed as a percentage, lower values indicate less residual variance. In our case a lower nMAE or nRMSE indicates a higher prediction accuracy on the dataset. The average of mean error is normalised over the total score range (s).

$$nMAE = \frac{\sum |f_t - a_t|}{s_{max} - s_{min}} \tag{3}$$

nRMSE The RMSE is defined as the square root of the mean square error. It is a standard way to measure the error of a model in predicting quantitative data.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \tag{4}$$

Though there is no consistent means of normalisation in the literature, our choice is to mean the range of the measured data, which is the most common choice in normalisation. We divide the RMSE by the score range (s).

$$nRMSE = \frac{RMSE}{s_{max} - s_{min}} \tag{5}$$

The ranges of values in the "annotations" columns in table 2 are defined as

(i) age: *years (int)*

(ii) gender: {*0,1*}

(iii) ethnic: *[0,1]*

(iv) height: *meters (float)*

(v) weight: *kilograms (float)*

(vi) sports: {*Alpine Skiing, Biathlon Bobsleigh, Cross Country, Curling, Figure Skating, Freestyle Skiing, Ice Hockey, Luge, Short Track, Skeleton, Ski Jumping, Snowboard, Speed Skating*}

(vii) glasses: {*0,1*}

(viii) attractiveness: *[0,1]*

(ix) complexion: {*fair, medium, olive, deep*}

(x) hair colour: {*blond, light brown, brown, dark brown, black*}

(xi) hair style: {*pony, short, long, occlusion eye, occlusion cheek*}

To measure a bias, we evaluate our network on our database of synthetically created images (GAN), called the Eurasian dataset. In this way we are able to classify the ethnicity exactly and evaluate the difference of aesthetic score, being our bias. The idea of this experiment is that in a fair network, ethnicity does not affect the aesthetic rating. In a fair network all lines should be at the same height for the same age class, as mentioned in fig. 6.

B Introduction to Ethics of AI

This section is a addition to the introduction of this work. While it does not provide information about the technical issue, it provides an introduction the ethics that drive the development of AestheticNet.

B.1 Moral Enhancement through AI

Faced with an increasingly diversified global environment, racial, religious, and cultural conflicts continue, and human morality is also facing unprecedented challenges. Many scholars have put forward the idea of human moral enhancement through technology, such as biotechnology, these interventions include the use of various substances, such as oxytocin and serotonin, as well as of various techniques, including transcranial magnetic stimulation and the provision of neurofeedback (Kelemen et al., 2015; Lara & Deckers, 2020).

The aim of these interventions would be to promote trust in others and to foster the desire to collaborate, but the harm that biotechnology does to the human body has always been a matter of concern. Many of us are biologically predisposed to have limited cognition and levels of altruism (Baron-Cohen, 2012; Wallace et al., 2007).

Everyone has an implicit bias, although this is not entirely caused by biological reasons. The existence of prejudice will lead to the problem of bias in machine learning, which has been proved in previous experiments. Compared with biotechnology, artificial intelligence is more trustworthy in helping humans with moral enhancement while it is improving the quality of human life. Especially existing implicit bias interventions tend to produce limited effects. For example, an air quality detection system installed on the road can remind us whether we should limit the number of private car trips and reduce gas emissions (Lai & J., 2016). An intelligent virus tracking APP reminds you that you might have come into contact with high-risk groups and that it is now necessary to consider home isolation. Some researchers are very pessimistic about the moral nature of humans while Dietrich is very optimistic about the possibilities of AI and believes that robots would have achieved that "Copernican turn" inaccessible to most humans by their biological conditioning (Dietrich, 2011). AI could monitor physical and environmental factors that affect moral decision-making, could identify and make agents aware of their biases, and could advise agents on the right course of action, based on the agent's moral values (Kelemen et al., 2015).

B.2 Social Problems caused by AI

Machine learning dominated by human training inevitably enables artificial intelligence to replicate human bias, including human implicit bias. Implicit bias is a bias that may be unconscious or uncontrollable. It exists in almost everyone, and the resulting social problems are endless. These subconscious thoughts are learned through our experience and are so deeply rooted that we may ignore them. Temporary unconscious bias may be the source of discriminatory and discriminatory practices. For example, even though Americans consider themselves unbiased when measuring unconscious stereotypes 90% of whites (Moule, 2009) and 50% of blacks associate negative characteristics with black images. From the beginning of childhood, white children and children of colour have liked white dolls (News, 2009).

In 2010, child psychologist and University of Chicago professor Margaret Beale Spencer (leading researcher in the field of child development) was hired by CNN as a consultant. Her team tested 133 children from schools with specific economic and demographic requirements. Tests have shown that white children have a high reaction rate to "white prejudice". They identify the colour of their skin as positive, while darker skin is identified as negative. Dr. Spencer said that even black children are prejudiced against whites, but far fewer than white children (Jill Billante et al., 2010).

According to a Reuters report in October 2018, Amazon's AI recruiting tool discriminates against women (Dastin, 2018). The research and development team at Amazon has been creating applications since 2014 to check the performance of applicants using the trained AI recruitment engine. Despite many years of experience and an exceptionally long maturation period, such obvious disadvantages are difficult or even impossible to recognise. Judging from the resumes sent to Amazon by the AI recruitment engine over the past ten years, most of them are men. The system automatically lowered the ranks of the two women's universities and associated the keyword "women" with "captain of the women's chess club". As one of the nine major AI companies that go hand-in-hand with Google and Facebook, the AI discrimination scandal that Amazon has fallen into has aroused considerable repercussions in the field of artificial intelligence and has aroused heated discussions from all walks of life.

On May 25, 2020, in Minnesota, USA, a black man died after being kneeled on by police for 7 minutes (Hill et al., 2020). In an interview with CBS, Minneapolis Mayor Jacob Fry said, "I don't know if there is explicit or implicit racism, but racism is definitely involved".

As we needed a larger dataset for image annotation, we use recently published generative adversarial networks (GAN), such as StarGAN-v2 (Choi et al., 2020). For our annotation process we create portraits of females and manually selected outcomes with no or only very few artefacts. In the annotation process, the annotators label a mix of GAN images and real images from the SCUT FBP (Liang et al., 2018) dataset. After we evaluate the results of the annotation, we realise that the highest annotations are dominated by images generated with StarGAN. With our dataset, we experience no uncanny valley problems. We can assume that during clicking the border between real person and synthetically generated person blurs. This is also demonstrated on the website *thispersondoesnotexist.com* with the accompanying publication by Karras et al. (Karras et al., 2019). In addition, generative networks producing synthetically generated portraits and animations, called deepfakes, are currently on the rise. Simultaneously, due to current circumstances, meetings using video-calls are currently on the rise. It could be argued that soon, within a video call, it will not possible to differentiate between a real human being or a deepfake of a person by vision alone. This creates many more questions on social problems which might be created by AI.

C WHAT IS UNCONSCIOUS BIAS?

Obviously, unconscious bias is a kind of prejudice that can be expressed as positive preference or negative discrimination. Favouritism is morally acceptable in many cases. A preference for both one's own inner group, such as family and friends, and for the support of a weak social group usually does not trigger the exclusion of external groups. The preference in this case did not lead to obvious social problems. However, if the prejudice encompasses a large, specific group, such as ethnicity, gender, or age, then moral and ethical problems are more likely to arise.

Another situation is that when the result of preference leads to unfair treatment of a certain individual or group, discrimination also occurs, and its counterpart is explicit bias. Inner group preference is a common cause of unconscious bias, but as explained earlier, it is also likely to lead to discrimination.

Unconscious bias is often used interchangeably with implicit bias in the fields of philosophy and psychology. Because unconscious bias is literally easier to understand and accept by the general public, it is used more often in a wider range of everyday language. Implicit bias was first defined by psychologists Mahzarin Banaji and Anthony Greenwald in 1995, where they argued that social behaviour is largely influenced by unconscious associations and judgements, corresponding to it is explicit bias.

However, most psychologists have abandoned Freud's psychoanalytic theory of unconscious mental processes. Unconscious bias usually manifests as a stereotype of things. It manifests in many forms and often occurs in our daily lives. The most common manifestations are prejudices and stereotypes

about affinity, gender, or the appearance of people. For example, the unconscious bias of the title professor is automatically assigned to the portrayal of an older man, although there are also many young and female professors. In addition, many people have the stereotype that women are worse at math problems and better at verbal problems than men (Johns et al., 2005).

Affinity bias refers to when you unconsciously prefer people who share qualities with you or someone you like. This bias was also verified in our first part of the experiment. Networks trained by Europeans think European faces are more beautiful than Asian faces, in contrast, networks trained by Asians think Asian faces are more beautiful than European faces. A study at Yale shows the "male" candidate was judged to be more talented and experienced; he was selected for the job more often and at a higher salary (Moss-Racusin et al., 2012).

Obviously, this is a gender bias. And when you unconsciously notice people's appearances and associate it with their personality, you might have beauty bias. Untidy appearance does not mean that this is a person who lacks the ability to manage themselves. A person in shabby clothes may be economical but he is not necessarily poor. Unconscious biases like these happen to everyone every day. And as proved later in our experiment, the bias in the training data is likely to be given to the artificial intelligence, which replicates the bias.

C.1 Causes of Unconscious Bias

1. Susceptibility to bias. We are used to the brain's fast, emotional, unconscious thinking mode. Kahneman states in (Kahneman, 2011) that there are two systems in the brain to organise our daily life: System 1 and System 2. Fast, emotional, and unconscious activities like driving, talking, or cleaning use System 1 since it requires little or even no effort, but it is often prone to errors. System 2 is slow, logical, effortful, conscious thought, where reason dominates. System 1 is a kind of mental shortcut, and we take this shortcut. Rules of thumb, educated guesses, and using "common sense" are all forms of mental shortcuts. Implicit bias is a result of taking one of these cognitive shortcuts inaccurately (Rynders, 2019). As a result, we incorrectly rely on these unconscious stereotypes to provide guidance in a very complex world. Especially when we are under high levels of stress we are more likely to rely on these biases than to examine all the relevant surrounding information (Wigboldus et al., 2004).

2. We seek patterns. One key reason we develop such biases is that our brains have a natural tendency to look for patterns and associations in order to make sense of a very complicated world. Research shows that even before kindergarten, children already use their group membership (e.g., racial group, gender group, age group, etc.) to guide inferences about the psychological and behavioural traits. At such a young age, they have already begun to seek out patterns and recognise what distinguishes them from other groups. Not only do children recognise what sets them apart from other groups, they believe "what is similar to me is good, and what is different from me is bad" (Cameron et al., 2001). Children aren't just noticing how similar or dissimilar they are to others, but also that dissimilar people are actively disliked (Aboud, 1989). Recognising what sets you apart from others and then forming negative opinions about those outgroups (a social group with which an individual does not identify) contributes to the development of implicit biases.

3. Social and cultural influences. Influences from media, culture, education, and your individual upbringing can also contribute to the rise of implicit associations that people form about the members of social outgroups. Media has become increasingly accessible, and while that has many benefits, it can also lead to implicit biases.

The way TV portrays individuals, or the language journal articles use, can ingrain specific biases in our mind. They can lead us to associate Black people as criminals or females as nurses or teachers. How children are raised can also play an important role. One research study found that parental racial attitudes can influence children's implicit prejudice (Sinclair et al., 2005). Parents are not the only figures who can influence such attitudes. Siblings, the school setting, and the culture in which you grow up can also play a role in shaping your explicit beliefs and implicit biases.

Social education also has a powerful effect because it includes not only traditional school education but also family education and self-study. Learning is the process of acquiring new understanding, knowledge, behaviour, skills, values, attitudes, and preferences (Gross, 2010).

From a blank sheet of paper at birth to receiving education and learning, gaining knowledge and gradually forming their values and understanding of surrounding things, there is no doubt that education and learning play a vital role in a person's will and thinking. A common belief is that education is an important determinant to racial prejudice, and there is preliminary evidence that the effect of this education varies from country to country (Hello et al., 2002).

Research scholars have found that the Polish education system plays a decisive role in the nationalism and prejudice of students (Żuk, 2018). In the medical field, because the attitude of medical staff to obese individuals has contributed to discrimination and led to poor health, the medical education environment may have explicit and implicit biases against obesity. Researchers who have adopted innovative educational interventions (read about obesity drama) found that it has a significant effect on reducing implicit prejudice against obese people (Matharu et al., 2014).

C.2 Influence on Artificial Intelligence

Numerous studies show that human-trained machines repeat human bias (Zuiderveen Borgesius, 2018; Stephens-Davidowitz, 2014; Munger, 2017; Horvitz, 2017). The assumption is to create unbiased artificial intelligence by inserting labelled data from people who are free of prejudice. Unfortunately, everyone is biased because a large part of human bias is unconscious and hard to detect. Unconscious prejudices affect 90 to 95 percent of people. Psychologists demonstrated this at a press conference at the University of Washington and presented a new tool that measures the unconscious roots of prejudice (Schwarz, 1998).

All science suffers from human bias. Even if we train giant robots to collect, store, and manipulate data for us, the ultimate observers, in the final analysis interpreters, and mediators of that data, are humans (Mukherjee, 2016).

Unconscious bias comes from the education background, the culture, attitudes, and stereotypes we pick up from the world we live in, and research over time and from different countries shows that it tends to line up with general social hierarchies. In words, unconscious bias is part of human nature and affects everyone, whether they realise it or not. Therefore, we can only try to avoid it and work with biased data; there is no way to completely eliminate it.

What is the significance of our research for the society? Firstly, from the field of image vision that we are engaged in, AI bias also exists in the field of image processing. Artificial intelligence applications tag pictures of White American brides as "brides", "dresses", and "weddings" while pictures of North Indian brides are tagged as "performing arts" and "costumes" (Shankar et al., 2017). Angwin's and Larson's (Larson & Angwin, 2016) analysis of ethical bias has prompted research showing that the disparity can be addressed if the algorithms focus on the fairness of outcomes. That which applies automatic labels to pictures in digital photo albums, was classifying images of black people as gorillas (BBC, 2015). Since the data set does not contain enough ethnic minorities, the artificial intelligence judges which designed by beauty. AI does not like black-skinned women (beauty.ai, 2016). Just imagine this would be a job position for a cover model position: there is no doubt that candidates with black skin will be rejected.

Although both Hume (David, 1898) and Kant (Kant & Guyer, 2000) believe that aesthetic judgement is only a subjective feeling which regarding the pleasure that we take from a beautiful object, and aesthetics itself is neither right nor wrong nor moral, the behaviour caused by aesthetic judgements is related to morality (Cui et al., 2019; Haidt, 2001). Many studies show that unfair recruitment cases are encountered due to aesthetic judgements (Maddox & Perry, 2018; Mason, 2017; Beattie & Johnson, 2012). More attractive people have higher incomes than less attractive individuals (Anýžová & Matějů, 2018; French, 2002; Cawley, 2004). More and more companies use AI enhanced recruitment systems, such as Facebook, LinkedIn and Unilever. The fairness of the image processing system is very important to the entire AI recruitment system. Our research provides a practical example of how to build a fair and unbiased AI.

Furthermore, through this research, we hope the unconscious bias will get more social attention by studying how unconscious bias affects decision-making in artificial intelligence. What is gratifying is that many companies are beginning to pay attention to the negatives that unconscious bias may entail and are making continuous efforts. Examples are Google (Google LLC, 2016) and Facebook (Facebook Inc., 2014). They are strengthening the training of employees in this area. Facebook

Table 3: Comparison of prediction accuracy on SCUT-FBP5500

	PC	nMAE (%)	nRMSE (%)
AlexNet (Liang et al., 2018)	0.8298	7.345	9.548
AlexNet (Zhai et al., 2019)	0.8634		
ResNet-18 (Liang et al., 2018)	0.8513	7.045	9.258
ResNeXt-50 (Liang et al., 2018)	0.8777	6.295	8.313
HMTNet(Xu et al., 2019)	0.8783	6.2525	
8.158 AaNet (Lin et al., 2019)	0.9055	5.590	7.385
P-AaNet (Lin et al., 2019)	0.8965	5.713	7.588
2M BeautyNet (Gan et al., 2020)	0.8996		
EfficientNetB3 based AestheticNet (ours)	0.9011	5.841	7.663
VGG-Face based AestheticNet (ours)	0.9363	4.400	6.261
AestheticNet (ours)	**0.9601**	**3.896**	**5.580**

designed a webpage to make unconscious bias training videos widely available and Google has put about 60,000 employees through a 90-minute unconscious bias training program. This will help reduce human bias during human-computer interaction, but the economic cost of the training is also huge. If there is an artificial intelligence similar to our research that can better circumvent human unconscious biases it would be a viable alternative.

D TRAINING

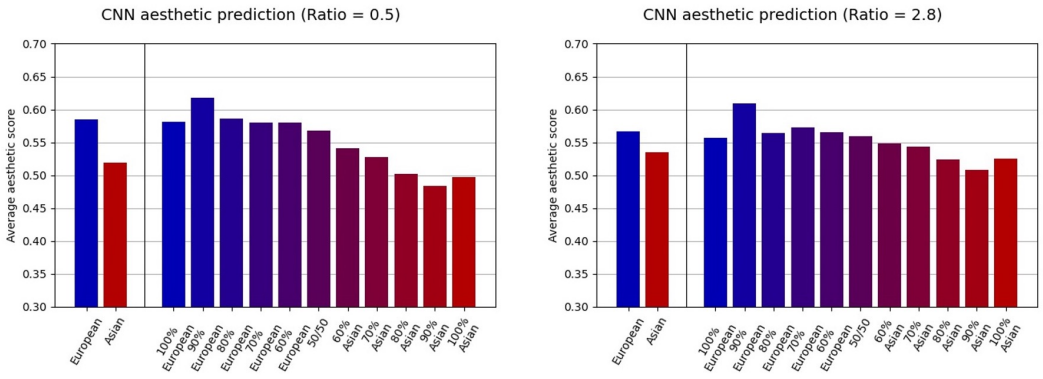

Figure 10: Effect of different ratios on the output of the network.

D.1 EVALUATION OF UNBIASED NETWORK

We compare our method with other state-of-the-art approaches on the SCUT-FBP500 datasets. As shown in table 3 our AestheticNet therefore significantly surpasses previous approaches, which is mainly due to the augmentation with synthetic images and the optimisation of the previous approaches. In our best experiment, we achieve a Pearson correlation of 0.9601, a normalised mean average error of 3.896% and a normalised root mean squared error of 5.580%. The results are normalised because there are different datasets with different score ranges.

Having a state-of-the-art aesthetic prediction network, we then train a third CNN on the features from the Asian and German labelled networks to generate a non-biased network. Therefore, the synthesised Eurasian dataset is used with a categorical-cross-entropy-loss-function to converge the subgroup's intersection points of fig. 6.

D.1.1 REMOVING BIAS USING CLUSTERED LABELS

A more sophisticated approach in getting rid of the bias in training data is our second approach. Within this we are developing a new method to reduce the bias in the training data. This method

consists of a deep learning network that is trained on the original learning task within the data set, and then minimises the bias inside the learned latent distributions using a specially adapted loss function.

Each data record contains a list of labels $a = a_1, ..., a_n$, which are to be debiased, and a further list of labels $b = b_1, ..., b_n$. In this example we remove the bias from the ethnical label a_1 and preserve the age, profession, hair colour and skin complexion labels. The network evaluates all attributes of the data set during the training and groups all objects according to the attributes b in clusters.

Within each subgroup the difference between the ethnical mean value $\overline{a_1}$ represents the bias. A nonlinear operation, similar to the gamma correction in image systems, is then applied to the ethnic label to preserve the range of the values and bring the differences closer together. These differences for all clusters are the measure of the loss function, which is implemented as categorical cross entropy loss and should be minimised during training. With this we present a universally adaptable method to make any network fairer according to given labels.