Person Perception Biases Exposed: Revisiting the First Impressions Dataset

Julio C. S. Jacques Junior
Universitat Oberta de Catalunya, Spain
Computer Vision Center, Spain
jsilveira@uoc.edu

Agata Lapedriza
Universitat Oberta de Catalunya, Spain
alapedriza@uoc.edu

Cristina Palmero
Universitat de Barcelona, Spain
Computer Vision Center, Spain
crpalmec7@alumnes.ub.edu

Xavier Baró
Universitat Oberta de Catalunya, Spain
Computer Vision Center, Spain
xbaro@uoc.edu

Sergio Escalera
Universitat de Barcelona, Spain
Computer Vision Center, Spain
sergio@maia.ub.es

Abstract

This work revisits the ChaLearn First Impressions database, annotated for personality perception using pairwise comparisons via crowdsourcing. We analyse for the first time the original pairwise annotations, and reveal existing person perception biases associated to perceived attributes like gender, ethnicity, age and face attractiveness. We show how person perception bias can influence data labelling of a subjective task, which has received little attention from the computer vision and machine learning communities by now. We further show that the mechanism used to convert pairwise annotations to continuous values may magnify the biases if no special treatment is considered. The findings of this study are relevant for the computer vision community that is still creating new datasets on subjective tasks, and using them for practical applications, ignoring these perceptual biases.

1. Introduction

Psychologists have long studied human personality, and throughout the years different theories have been proposed to categorise, explain and understand it. From the past few years, it has also become an attractive research area in visual computing [39, 17], motivated by the fact that automatic methods for personality recognition or perception can be applied in a vast number of scenarios. Nevertheless, while real personality can be accessed through self-report questionnaires, perceived (or apparent) personality assessment is given by external observers through impression formation, and here is where person perception bias comes in.

Technologies for human behaviour analysis have shown their vulnerability to human annotation biases [12]. In particular, human bias is very strong when trying to infer personality attributes of someone during a first short encounter. This subjectivity makes the task of creating automatic personality perception systems challenging, since the biases will be reflected on the annotations and, consequently, on the resulting recognition systems. Therefore, creating methods that preserve human bias can have negative consequences if they are used in applications that deal with human outcomes. While the use of pairwise instance comparison [27, 6, 20] significantly reduces perception bias produced by absolute annotations, completely eliminating it in subjective tasks is extremely difficult.

This work uses the First Impressions (FI) [27] dataset to expose the existence of person perception bias in data labelling of personality. The FI dataset is one of the biggest publicly available datasets on the topic. Our work is based on recent studies that demonstrate the bias produced by perceived gender, attractiveness and age [28] during the impression formation. In particular, we derive perception biases from pairwise annotations and associated person's attributes[1]. For example, we show that women are more fre-

[1] Attribute categories used in this research are imperfect for many reasons. For example, there is no gold standard for "ethnicity" categories, and it is unclear how many gender categories should be stipulated (or whether they should be treated as discrete categories at all). This work is based on

quently perceived as more *Open to experience* than men, that older men are more frequently perceived as more *Conscientious* than younger ones, and that ethnicity has stronger influence than gender if African-Americans are compared to either Asians or Caucasians, which bring to light some annotators' bias. Fig. 1 illustrates how person perception can influence data labelling when a subjective task like personality perception is considered.

Supervised learning methods developed to recognise apparent personality from images or videos [17] require a label for each individual in the train data, and pairwise annotations are in general not used. For this reason, the pairwise annotations of the FI [27] dataset were originally converted to continuous values using [6]. Our study also reveals that the mechanism used to convert pairwise annotations to continuous values may magnify the biases, making stereotyping stronger. Finally, it is important to note that previous works (e.g., [17, 12, 41]) using the FI dataset are based on the continuous values originated from the pairwise labels, and this is the first time the original (raw) pairwise labels are analysed.

2. Ethical Implications[2]

Personality perception and its applications. People spontaneously build first impressions of unacquainted individuals in milliseconds, even from a still photograph, quite consistently [37]. However, such snap judgements, which are built and used to interact with others, are often stereotyped [30]. Therefore, do we want machines to do the same? Having machines that form first impressions of others has risks. Such systems are trained from human annotations and inherit human perception bias along with other biases created by culture, beliefs or previous experiences. Since it is highly likely that automatic personality perception is not accurate, these technologies are not ready to be used for legal applications or for anything that determines opportunities for people, such as job interviews. Furthermore, having access to people's personality (either if real or apparent) just by extracting and analysing data from any kind of input could represent a major threat to their privacy. Not only in terms of rights, but also because it could pave the way for effective mass manipulation and psychological persuasion [24]. On the other side, automatic personality perception can be very useful in social robotics [4], to design machines that can approach people in a natural way, creating more comfortable experiences and building trust [21]. In particular, applications related to health care, education or human assistance can benefit from using automatic impression formation.

Bias in face attributes recognition. Our work partially relies on automatic face classification methods to extract an attractiveness score and to estimate apparent age. Both methods suffer from the same type of perception bias previously described. According to [35], facial cues often guide first impressions and these first impressions guide our decisions. Face attractiveness, however, is very subjective and may be subject to critics when applied to social computing. Nevertheless, the topic has been widely studied in psychology/sociology [36, 23, 43, 26, 35]. These attributes have been selected to give visibility to the existing biases, especially because the well known "*attractiveness halo effect*" (i.e., more positive impressions are given to more attractive people) has its particular influence on data labelling.

3. Related Work

Fairness in machine learning [3] is rapidly gaining interest among the research community and industry. This has been partially motivated by the biased results reported in the literature (e.g., [45, 2, 12, 28]), along with the difficulty to interpret *latent* representations [29]. According to [14], fairness-aware machine learning approaches can be categorised as: 1) preprocessing techniques which aim to modify the input data; 2) algorithm modification techniques, which modify existing algorithms by adding constraints or regularisation; and 3) postprocessing techniques which modify the output of an existing method to be fair. These categories consider the data is already available and ready to use. Our work, however, goes one step back and analyses how perception bias affects data labelling of a subjective task, which aligns with the idea that unfairness induced by unmeasured predictive variables should be addressed through data collection [7]. Thus, rather than addressing general bias problems such as imbalanced training data, covariate shift or sample selection [15], which can be found in almost any machine learning-based task, this work focuses on the biases coming from *human perception*.

In visual perception, contextual effects and prior experience lead to systematic biases in the judgement [8]. Cognitive and perceptual biases have distinct causes and effects, and can be grouped into different categories [10] given the bias type (e.g., fundamental attribution error, cultural bias, belief bias, selective perception, among others). The biases produced by human perception, which have been widely studied in sociology and psychology (e.g. [25, 35]), have a strong influence in subjective tasks such as automatic personality perception [17], (job) recommendation systems [12], emotion recognition [32] or image captioning [2]. However, works from a psychological perspective are limited to perform statistical analysis on small-scale datasets. On the other hand, most works from a computational perspective [38, 5, 45, 19, 40, 42, 34] study the general bias problems [15] mentioned above, while little attention is given to subjective bias analysis [32, 29, 31, 41] be-

an ethical and legal setting, and the methodology and findings are expected to be applied later to any re-defined and/or extended attribute category.

[2]For more information about ethics in AI, we refer the reader to [13].

Figure 1. Imagine that pairs of short videos are given, with people talking to the camera about any predefined topic. As annotator, you are asked to define what individual in a pair looks more friendly, more organised or maybe more authentic. Then, you may start analysing people's behaviour and attributes in order to build your first impressions. At the end, your choices might tell something about you. However, the overall perception given hundreds or thousands annotators may tell something about the database. Snapshots from the First Impressions [27] dataset (attributes empirically defined for illustration purposes).

yond the perspective of explainable models [11, 16, 28, 12].

In [32], authors show that the order of how images are displayed to the annotators may significantly bias the labels in facial emotion recognition tasks, whereas [29] proposed a data-to-data translation approach by learning a mapping from an input domain to a fair target domain, where a fairness constraint is enforced. The latter focused on analysing the gender attribute and the overall goal was to maximise equal opportunity between males and females. Robinson et al. [31] showed that the performance gaps in face recognition can be reduced by learning subgroup-specific thresholds, revealing that the conventional approach of learning a global threshold may also bias the results. More recently, Yan et al. [41] investigated the biases on multimodal systems designed for automatic personality perception, using the FI [27] dataset as case study. The study revealed that different modalities show various patterns of biases, and that data fusion also introduces additional biases to the model. Thus, they propose two debiasing approaches based on data balancing and adversarial learning to mitigate the biases. The analyses performed in their work, however, are based on the continuous values provided with the FI [27] dataset, and the original pairwise annotations are not considered.

Collecting labels for subjective tasks is challenging. Biased annotations are particularly difficult to detect and correct. For annotation tasks related to subjective human behaviour and personality attributes [17], pairwise comparison is becoming a standard procedure, as it has demonstrated [27, 6, 20] to be very effective at mitigating labeller biases. For instance, Joo et al. [20] asked Amazon Mechanical Turk workers to compare a pair of images in face trait dimensions rather than evaluating each image individually. A similar strategy was applied in [27, 11, 12] for video files.

Comparison schemes have three main advantages in data labelling for person perception: 1) they naturally identify the strength of each sample in the context of relational distance from other examples, generating a more reliable ranking of subtle signal differences [20]; 2) they mitigate the sequential bias [32], e.g., scoring someone very low on a certain dimension because of an unconscious comparison with previous samples where the score was high; and 3) the annotators do not need to establish the absolute baseline or scales for any dimension, which would be unnatural. Although pairwise ratings significantly reduce the bias in person perception annotation tasks, this work shows that people's attributes, combined with annotators' bias, can have a strong influence on data labelling. This suggests that future works on the topic need to pay attention to the way the pairs are defined and presented to the annotators, since the pairs themselves can also be a source of bias, particularly for sensitive applications where reducing biases under certain controlled dimensions is crucial.

4. The First Impressions Dataset

The ChaLearn First Impressions (FI) dataset [27] is currently the largest, public and labelled dataset developed to advance research on automatic personality perception. The FI dataset was released in the context of a computational challenge, where the goal was to automatically recognise the Big-Five (OCEAN) apparent personality traits of single individuals in videos: ***O**penness to experience*, ***C**onscientiousness*, ***E**xtraversion*, ***A**greeableness*, and ***N**euroticism*[3]. Later, it was labelled with an "Invite to interview" variable, aiming to advance research on explainable machine learning [11]. The dataset is composed of 10K short video clips (average duration of 15s each) extracted from more than 3K different YouTube videos of people talking to a camera. Some snapshots of the dataset are shown in Fig. 1, while Fig. 2 shows the pairwise-based annotation interface. The database was annotated using crowdsourcing, being each pair annotated by one single annotator. In this work, we release[4] and analyse by the first time the original pairwise annotations of the First Impressions dataset.

Gender and ethnicity labels are also provided with the data (both provided via crowdsourcing, i.e., they are perceived attributes). Overall, the dataset is more or less bal-

[3] *Neuroticism* was labelled in [27] as *"Emotion stability"*, which is the opposite of Neuroticism. This will be represented along the paper as $\overline{\text{N}}$.

[4] The pairwise annotations of the FI dataset can be found at `http://chalearnlap.cvc.uab.es/dataset/24/description/`.

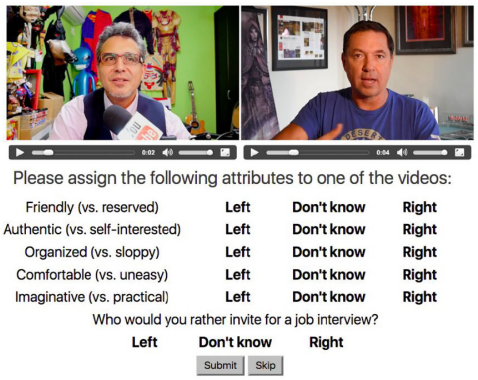

Figure 2. The interface used for pairwise data labelling [27].

anced in gender (45% males and 55% females). However, it is very imbalanced in terms of ethnicity (3% Asian, 86% Caucasian and 11% African-American).

The dataset has around 345K video pairs labelled with apparent Big-Five personality traits and the "Interview" variable. However, some pairs were labelled with the "*Don't know*" label (illustrated in Fig. 2) for some dimensions, for which the annotators were not so confident about the ranking. Table 1 shows the number of "valid pairs" (i.e., when ignoring the "*Don't know*" label) per dimension, as well as for different subsets given the gender/ethnicity of individuals being compared. As it can be seen, data imbalance is strong with respect to the different subsets, which imposes another obstacle in addition to perception bias when the goal is to build fair machine learning methods.

Table 1. Number of "valid pairs" per trait and per subset, given the gender/ethnicity of individuals in a pair.

	O	**C**	**E**	**A**	**N̄**	**Interview**
Valid pairs	307513	313749	321684	318792	321078	323178
Per Gender						
Male vs. Female	152365	155466	159467	157829	158942	160095
Female vs. Female	91931	93795	96231	95483	96080	96711
Male vs. Male	63217	64488	65986	65480	66056	66372
Per Ethnicity						
Cauc. vs. Cauc.	227558	232160	238080	235944	237546	239142
Afr-Am. vs. Cauc.	56496	57586	59100	58543	59008	59387
Asian vs. Cauc.	17367	17770	18137	18011	18155	18264
Afr-Am. vs. Afr-Am.	3558	3642	3694	3654	3710	3702
Asian vs. Afr.	2204	2261	2330	2296	2311	2341
Asian vs. Asian	330	330	343	344	348	342

Having the data labelled through pairwise comparisons, the pairwise data is converted in [27, 11] to continuous values using [6]. This method individually converts the ordinal ratings of each dimension into continuous values (such as the level of "*Extraversion*") by fitting a Bradley-Terry-Luce (BTL) model with maximum likelihood, which are further scaled to be in the range of $[0, 1]$. This way, each video sample in the dataset will have a continuous value associated to each trait dimension, which can be used by any supervised learning method, in a classification or regression task.

5. Automatic extraction of face attributes

This section describes how face attractiveness and perceived age of people present in the FI dataset are obtained. To remove any bias caused by the imbalanced ethnicity category, only Caucasian individuals are considered.

First, a face detector [44] is applied on each video at 5 consecutive frames. Then, face attributes are extracted using a modified version of the VGG-16 [33] model, that regresses either the attractiveness score or the perceived age, depending on the given task. Finally, the per-frame predictions are averaged per attribute. The proposed modification consists in removing the last layers of the original VGG-16 model (illustrated in Fig. 3 by a red box) and the inclusion of a convolutional layer (to reduce dimensionality) and three fully connected (FC) layers to learn hidden representations (using ReLu as activation function) before a final Dense layer (with *Sigmoid* activation) responsible for regressing the face attribute.

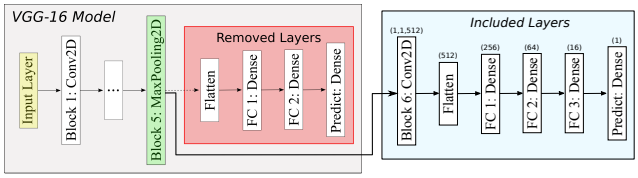

Figure 3. Modified VGG-16 model used to predict either face attractiveness or perceived age (depending on the given task).

Fig. 4 shows the distribution of face attributes extracted for all Caucasian individuals in the FI dataset. It must be emphasised that the aim of our work is not to advance the state of the art on face attribute recognition. Predicted attributes are taken as "truth" (i.e., soft labels, more precisely) due to the low error rates obtained on the associated datasets, detailed next, and used as proof of concept.

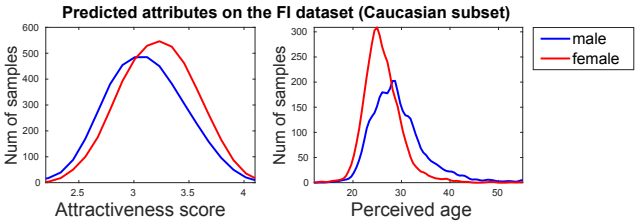

Figure 4. Distributions of predicted face attributes.

5.1. Face attractiveness

To recognise the attractiveness score of each individual, our model was trained with the SCUT-5500 [22] database. This dataset consists of 5.5K frontal unoccluded faces, with neutral expression, aged from 15 to 60 years old. It contains 4K images of Asians and 1.5K images of Caucasians, equally distributed in gender for each set. Images were labelled with beauty scores in the range of $[1, 5]$ by a total

of 60 volunteers aged from 18-27, which is also subject to impact the ground truth due to their implicit bias [43].

To evaluate the effectiveness of our model to predict attractiveness score on the SCUT-5500 [22] database, 85% of the data was randomly selected for training and the remaining samples for testing. Obtained Mean Absolute Error (MAE) on the test set was 0.247, comparable to results obtained in [22]. Note that we have also evaluated our model following the same protocol described above but considering Caucasian individuals only, however, obtaining a slightly higher MAE, most probably due to difficulty to generalise given the small-sized training data.

5.2. Perceived age

To automatically recognise the perceived age of each individual, the APPA-REAL [1] database was chosen. The database is composed of almost 8K images mainly showing a single person in frontal face, labelled with real and apparent age (in the latter case, via crowdsourcing), ranging from 0 to 95 years old. Our study, however, uses only the perceived age label as our intention is to analyse how the perception of age can bias pairwise data labelling. On average, each image was annotated with apparent age by 38 annotators, resulting in a very stable average apparent age (0.3 standard error of the mean).

To evaluate the performance of our VGG16-based model to predict apparent age, we followed the evaluation protocol defined in [1]. Obtained MAE on the test set was 7.12 (years), which is similar to results obtained by [18].

5.3. Training strategy

The two face attribute recognition tasks are trained in two stages. First, the model is initialised with weights pretrained on ImageNet [9]. Then, we train only the new layers. In a second stage, we fine-tune the whole model. Adam algorithm is used as optimiser, with learning rate $1\mathrm{e}{-5}$. Mean Squared Error is used as the loss function. Early stopping is performed if no decrease in validation error is observed. Finally, the best model for each task is kept based on the accuracy computed on the validation set.

6. Revealing the perception biases

This section reveals different perception biases found in the FI [27] dataset, from a global to a fine-grained analysis. In Sec. 6.1 and Sec. 6.2, we analyse perception biases associated to gender and ethnicity obtained directly from the pairwise (binary) labels, and show that some of them are amplified when converted to continuous values using [6]. To measure the perception bias present in the continuous values, we simply computed the number of cases where "individual A" obtained a higher continuous value than "individual B", given a particular trait/dimension and subset being analysed. The analyses consist of comparing subsets of data composed of pairs of individuals with particular attributes, e.g., "Male vs. Female" or "Asian vs. Caucasian", to show how some groups were perceived differently, in some cases, as a function of their attributes. In Sec. 6.3, we analyse how facial attributes (i.e., face attractiveness and perceived age) influenced data labelling of the FI dataset.

6.1. Gender bias

Table 2 shows the percentage of individuals perceived as being a more/less representative sample for a particular trait/dimension considering the gender attribute only, obtained directly from the pairwise labels ("PL"), and given the continuous values provided with the FI dataset ("CV") for the same pairs[5]. The percentages shown in Table 2 are obtained from the subset of "valid pairs" where individuals being compared have different gender. As it can be seen, there is an overall bias towards women, which is stronger for some traits (e.g., "O", "C" and "E"). Interestingly, the bias is amplified for all variables during the conversion from pairwise data to continuous values using [6]. Therefore, some traits are more impacted than others.

Table 2. Gender bias ("**M**ale vs. **F**emale" subset), measured on the pairwise labels (PL) and continuous values (CV) provided with the FI dataset. It can be observed an overall bias towards women. Moreover, differences are amplified when data is converted from pairwise labels to continuous values.

	O		**C**		**E**		**A**		**$\overline{\mathbf{N}}$**		**Interview**	
	M	F	M	F	M	F	M	F	M	F	M	F
PL	46.4	53.6	47.9	52.1	**44.7**	**55.3**	50.3	49.7	48.6	51.4	48.2	51.8
CV	**38.4**	**59.6**	**43.8**	**54.5**	**36.7**	**61.6**	49.3	48.4	45.4	52.7	44.7	53.6

6.2. Ethnicity and gender biases

Table 3 shows that gender had stronger influence than ethnicity when "Asian vs. Caucasian" subset is considered, and that there is an overall bias towards women, which is evidenced when pairs composed of individuals of different gender are used. On the contrary, Table 4 and Table 5 show that ethnicity had stronger influence than gender when subsets "Asian vs. African-American" and "African-American vs. Caucasian" are used. In these cases, Asians and Caucasians were more frequently perceived as being a more representative sample for a particular trait, compared to African-Americans, independently from the gender of the individuals. We can also observe a significantly lower number of pairs where both individuals are male, compared to other cases (especially in Table 4), which may also bias the analysis. Furthermore, as observed when analysing Table 2, some biases were magnified when converting the binary labels to continuous values using [6]. As expected, it seems

[5]The subset of individuals perceived as more/less representative sample for a trait is shown in tones of red/blue, respectively (from Table 2 to Table 5). Differences $\geq 10\%$ are shown in bold. Note, the "CV" sum may not be 100% as some pairs received the same continuous value.

the biases are amplified during the conversion from "PL" to "CV" as a function of the bias in "PL", i.e., the higher the bias in pairwise labels, the higher will be the magnification when converted to continuous value. This effect showed to be stronger for smaller subsets.

Table 3. Ethnicity and gender bias ("**Asi**an vs. **Cau**casian" set), measured on the pairwise labels (PL) and continuous values (CV) provided with the FI dataset. In this case, gender showed a stronger influence than ethnicity (towards women).

	O		**C**		**E**		**A**		**$\overline{\text{N}}$**		**Interview**	
	Global											
	Asi	***Cau***	***Asi***	***Cau***	***Asi***	***Cau***	***Asi***	***Cau***	***Asi***	***Cau***	***Asi***	***Cau***
PL	50.4	49.6	50.9	49.1	52.7	47.3	49.9	50.1	50.0	50.0	50.9	49.1
CV	49.2	48.3	50.7	47.4	**55.0**	**43.4**	47.8	50.0	47.9	50.1	49.8	48.3
	Male vs. Male											
Tot.	2431		2495		2522		2504		2517		2549	
PL	50.2	49.8	51.1	48.9	54.8	45.2	50.6	49.4	51.6	48.4	51.1	48.9
CV	48.7	48.5	51.7	46.2	**59.1**	**39.2**	48.6	49.1	50.1	47.8	50.8	47.0
	Female vs. Female											
Tot.	6459		6633		6756		6730		6771		6810	
PL	48.9	51.1	50.7	49.3	51.5	48.5	50.8	49.2	49.1	50.9	50.9	49.1
CV	46.7	50.9	48.6	49.5	50.7	47.6	48.5	49.5	46.6	51.4	48.5	49.6
	Asian Male vs. Caucasian Female											
Tot.	2561		2619		2697		2656		2707		2724	
PL	**44.9**	**55.1**	46.7	53.3	**44.5**	**55.5**	47.6	52.4	47.4	52.6	46.7	53.3
CV	**32.7**	**64.5**	**40.8**	**57.4**	**38.1**	**60.4**	44.4	53.5	**40.9**	**56.6**	**39.7**	**58.2**
	Asian Female vs. Caucasian Male											
Tot.	5916		6023		6162		6121		6160		6181	
PL	54.6	45.4	52.7	47.3	**56.6**	**43.4**	49.5	50.5	51.6	48.4	52.7	47.3
CV	**59.3**	**38.3**	**57.1**	**41.3**	**65.4**	**33.0**	48.2	49.5	51.6	46.6	**55.3**	**43.0**

Table 4. Ethnicity and gender bias ("**Asi**an vs. **Afr**ican-American" subset), measured on the pairwise labels (PL) and continuous values (CV) provided with the FI dataset. In this case, ethnicity showed a stronger influence than gender, i.e., there is an overall bias towards Asian individuals.

	O		**C**		**E**		**A**		**$\overline{\text{N}}$**		**Interview**	
	Global											
	Asi	***Afr***	***Asi***	***Afr***	***Asi***	***Afr***	***Asi***	***Afr***	***Asi***	***Afr***	***Asi***	***Afr***
PL	54.4	45.6	53.6	46.4	54.8	45.2	52.8	47.2	51.5	48.5	53.3	46.7
CV	**60.1**	**37.3**	**61.2**	**36.7**	**63.6**	**34.9**	**57.2**	**40.3**	**56.3**	**41.7**	**60.9**	**37.2**
	Male vs. Male											
Tot.	187		191		196		194		197		204	
PL	51.3	48.7	**55.5**	**44.5**	54.1	45.9	51.5	48.5	49.7	50.3	52.5	47.5
CV	51.9	46.0	**60.2**	**39.3**	**61.2**	**37.2**	52.6	44.3	48.7	46.2	**54.4**	**43.6**
	Female vs. Female											
Tot.	1051		1077		1124		1096		1113		1118	
PL	54.5	45.5	54.2	45.8	54.0	46.0	52.6	47.4	52.3	47.7	53.8	46.2
CV	**64.5**	**32.7**	**63.1**	**34.4**	**66.6**	**32.0**	**61.0**	**36.1**	**61.0**	**37.8**	**64.7**	**33.3**
	Asian Male vs. African-American Female											
Tot.	428		439		460		457		451		470	
PL	**56.5**	**43.5**	**56.5**	**43.5**	**57.2**	**42.8**	**57.3**	**42.7**	52.8	47.2	**56.0**	**44.0**
CV	48.6	48.8	**58.3**	**39.9**	**54.6**	**43.7**	**58.4**	**39.4**	**55.2**	**42.1**	**56.6**	**41.1**
	Asian Female vs. African-American Male											
Tot.	538		554		550		549		550		549	
PL	53.4	46.6	49.5	50.5	54.7	45.3	49.9	50.1	49.4	50.6	50.3	49.7
CV	**63.6**	**34.0**	**60.1**	**37.9**	**65.6**	**32.7**	50.1	47.9	50.4	47.6	**59.2**	**39.5**

6.3. Face attributes and related biases

Given the face attributes extracted for all Caucasian individuals (detailed in Sec. 5), we are able to analyse their influence on data labelling of the FI dataset. To remove the gender variable from the analysis, only pairs of individuals having the same gender are considered. For the sake of illustration, Fig. 5 shows the distributions of face attribute differences between pairs of individuals in this subset.

Table 5. Ethnicity and gender bias ("**Afr**ican-American vs. **Cau**casian" subset), measured on the pairwise labels (PL) and continuous values (CV) provided with the FI dataset. In this case, ethnicity showed a stronger influence than gender, i.e., there is an overall bias towards Caucasian individuals.

	O		**C**		**E**		**A**		**$\overline{\text{N}}$**		**Interview**	
	Global											
	Afr	***Cau***	***Afr***	***Cau***	***Afr***	***Cau***	***Afr***	***Cau***	***Afr***	***Cau***	***Afr***	***Cau***
PL	46.4	53.6	46.8	53.2	47.1	52.9	47.6	52.4	47.8	52.2	47.0	53.0
CV	**38.7**	**59.1**	**41.7**	**56.6**	**42.5**	**55.9**	**41.0**	**56.6**	**42.7**	**55.5**	**41.4**	**56.8**
	Male vs. Male											
Tot.	8364		8559		8726		8655		8737		8792	
PL	47.7	52.3	48.7	51.3	48.7	51.3	49.1	50.9	49.6	50.4	48.8	51.2
CV	**43.5**	**54.3**	46.5	51.9	45.9	52.3	47.1	50.5	49.1	49.1	47.0	51.3
	Female vs. Female											
Tot.	20339		20667		21254		21050		21221		21347	
PL	**44.7**	**55.3**	**45.4**	**54.6**	**44.9**	**55.1**	46.9	53.1	46.6	53.4	45.6	54.4
CV	**33.8**	**64.3**	**37.7**	**60.5**	**37.4**	**61.0**	**38.4**	**59.3**	**38.7**	**59.6**	**37.7**	**60.5**
	African-American Male vs. Caucasian Female											
Tot.	9411		9594		9865		9792		9853		9915	
PL	**44.3**	**55.7**	46.5	53.5	**43.1**	**56.9**	49.9	50.1	48.7	51.3	47.4	52.6
CV	**32.3**	**65.6**	**40.4**	**57.8**	**33.8**	**64.8**	46.2	51.5	44.3	53.8	**41.6**	**56.8**
	African-American Female vs. Caucasian Male											
Tot.	18382		18766		19255		19046		19197		19333	
PL	48.6	51.4	47.7	52.3	50.9	49.1	46.6	53.4	47.7	52.3	47.5	52.5
CV	45.4	52.3	44.5	53.7	50.9	47.3	**38.5**	**59.1**	**43.2**	**54.8**	**42.9**	**55.2**

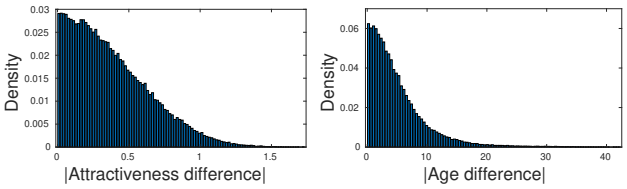

Figure 5. Distributions of face attribute differences between individuals in a pair ("Caucasian vs. Caucasian" subset, pairs composed of individuals of same gender).

6.3.1 Face attractiveness bias

Fig. 6 shows the number of times an individual in a pair recognised as "more attractive" was chosen, divided by the number of times an individual recognised as "less attractive" was selected, varying the face attractiveness difference between them. If face attractiveness had no influence, the ratio would be 0.5, meaning that individuals recognised as more or less attractive were rated equally. Values higher than 0.5 show a bias towards "more attractive" individuals, whereas values lower than 0.5 a bias towards "less attractive" ones. As it can be seen, when face attractiveness difference between individuals increases, there is a higher fraction of individuals in the pairwise data recognised as "more attractive" being perceived as a more representative sample for a particular trait, suggesting that face attractiveness is biasing the annotations. This trend was observed to be stronger for some traits, and even stronger when both individuals being observed are women. Therefore, as face

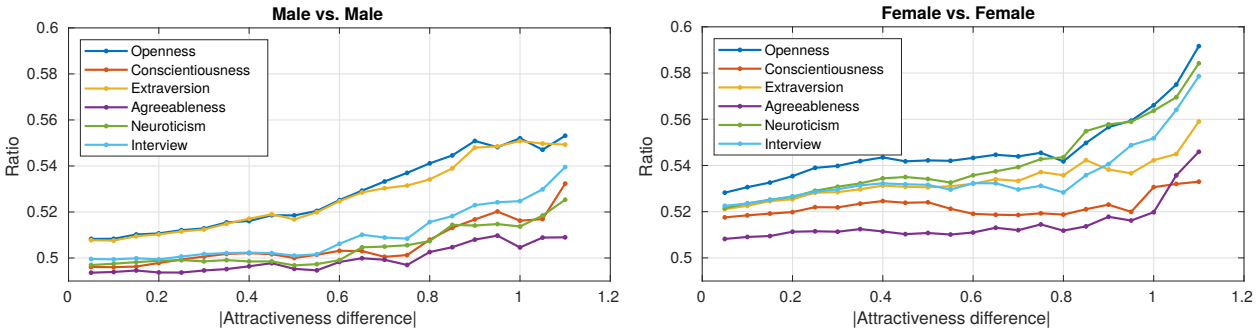

Figure 6. Face attractiveness bias ("Caucasian vs. Caucasian" subset). Number of times an individual in a pair recognised as "more attractive" was chosen, divided by the number of times a "less attractive" individual was selected, as a function of the attractiveness difference between them. Ratio > 0.5 indicates that "more attractive" individuals are more frequently selected, a trend that can be clearly seen from these plots, especially for larger differences. Note that in this case, *Neuroticism* relates to "Emotion stability".

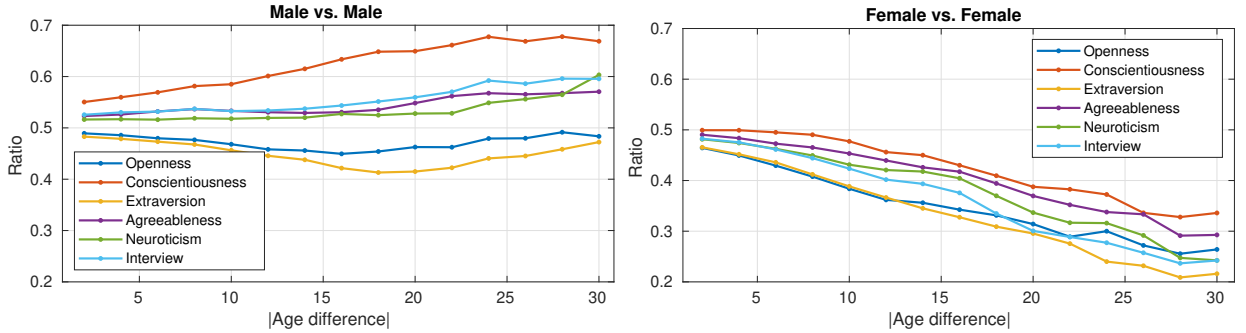

Figure 7. Perceived age bias ("Caucasian vs. Caucasian" subset). Number of times an individual in a pair recognised as "older" was chosen, divided by the number of times a "younger" individual was selected, as a function of the perceived age difference between them. Ratio values > 0.5 indicates that "older" individuals are more frequently selected, while ratio < 0.5 the opposite, i.e., "youngers" are more frequently chosen. The plots show a overall bias towards older man and a clear bias towards younger women. Note, in this case, *Neuroticism* relates to "Emotion stability".

attractiveness difference between individuals increases, the number of pairs being analysed decreases (Fig. 5, left image), which may affect the analysis, particularly for the cases where large differences are observed.

6.3.2 Perceived age bias

Fig. 7 shows the number of times an individual in a pair perceived as "older" was chosen, divided the number of times an individual perceived as "younger" was selected, varying the perceived age difference between them. If perceived age had no influence, the ratio would be 0.5, meaning that individuals recognised as older or younger were equally perceived. Values higher than 0.5 show a bias towards older individuals, whereas values lower than 0.5 a bias towards youngers. As it can be seen, annotators had an overall bias towards older men (except for traits "O" and "E"), especially when age difference between individuals in a pair increases, and a bias towards younger women most of the time for all dimensions, suggesting that the perceived age attribute influenced data labelling in different ways. Therefore, as perceived age difference between individuals increases, the number of pairs being analysed decreases (see Fig. 5, image on the right), which may affect the analysis, particularly for the cases where large differences are observed.

7. Final Considerations

This work used the First Impressions dataset as case study to expose how person perception can influence data labelling of a subjective task like personality. We analysed by the first time the original pairwise binary annotations provided with the FI dataset, and revealed the existence of different types of perception bias. This study also showed that the mechanism used to convert pairwise annotations to continuous values may magnify the biases if no special treatment is considered. Thus, this crucial step should be carefully revised, and possible negative consequences mitigated. In addition to gender and ethnicity biases, we analysed how the *attractiveness halo effect* and the perception of age can affect data labelling of personality, derived from the

pairwise annotations and face attributes automatically extracted. Although these perception biases have been widely studied in psychology and social sciences, the topic has received almost no attention from the computer vision community.

After analysing the pairwise-based annotation setup of the FI dataset, our study suggests that new protocols need to pay more attention to the way the pairs are defined and presented to the annotators, since the pairs themselves can be a source of bias. Moreover, as perception is dependent on the observer, the analysis and correlation of attributes between annotators and people being annotated could explain how some biases are produced. However, this would require a dedicated discussion around privacy and ethical issues that goes beyond the scope of this work.

Acknowledgment

This research was supported by Spanish projects TIN2015-66951-C2-2-R, RTI2018-095232-B-C22, and PID2019-105093GB-I00 (MINECO/FEDER, UE) and CERCA Programme/Generalitat de Catalunya. This work is partially supported by ICREA under the ICREA Academia programme.

References

[1] Eirikur Agustsson, Radu Timofte, Sergio Escalera, Xavier Baro, Isabelle Guyon, and Rasmus Rothe. Apparent and real age estimation in still images with deep residual regressors on appa-real database. In *International Conference on Automatic Face & Gesture Recognition (FG)*, pages 87–94, 2017.

[2] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[3] Sarah Bird, Krishnaram Kenthapadi, Emre Kiciman, and Margaret Mitchell. Fairness-aware machine learning: Practical challenges and lessons learned. In *Int. Conference on Web Search and Data Mining*, pages 834–835, 2019.

[4] Cynthia Breazeal. Social interactions in HRI: the robot view. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 34(2):181–186, 2004.

[5] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, volume 81, pages 77–91, 2018.

[6] Baiyu Chen, Sergio Escalera, Isabelle Guyon, Víctor Ponce-López, Nihar Shah, and Marc Oliu. Overcoming calibration problems in pattern labeling with pairwise ratings: Application to personality traits. In *European Conference on Computer Vision Workshop (ECCVW)*, pages 419–432, 2016.

[7] Irene Chen, Fredrik D Johansson, and David Sontag. Why is my classifier discriminatory? In *Advances in Neural Information Processing Systems*, pages 3539–3550, 2018.

[8] Ron Dekel and Dov Sagi. Perceptual bias is reduced with longer reaction times during visual discrimination. *Communications Biology*, 3:59, 2020.

[9] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

[10] E. Dimara, S. Franconeri, C. Plaisant, A. Bezerianos, and P. Dragicevic. A task-based taxonomy of cognitive biases for information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 26(2):1413–1432, 2020.

[11] H. J. Escalante, I. Guyon, S. Escalera, J. C. S. Jacques Junior, M. Madadi, X. Baró, S. Ayache, E. Viegas, Y. Güçlütürk, U. Güçlü, M. A. J. van Gerven, and R. van Lier. Design of an explainable machine learning challenge for video interviews. In *International Joint Conference on Neural Networks (IJCNN)*, pages 3688–3695, 2017.

[12] H. J. Escalante, H. Kaya, A. Salah, S. Escalera, Y. Güç;lütürk, U. Güçlü, X. Baró, I. Guyon, J. C. S. Jacques Junior, M. Madadi, S. Ayache, E. Viegas, F. Gurpinar, A. S. Wicaksana, C. Liem, M. A. J. Van Gerven, and R. Van Lier. Modeling, recognizing, and explaining apparent personality from videos. *IEEE Transactions on Affective Computing*, 2020.

[13] European Commission. Ethics guidelines for trustworthy ai. [online] Available at: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai [Accessed 5 Oct. 2020], Apr 2019.

[14] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 329–338. ACM, 2019.

[15] Jindong Gu and Daniela Oelke. Understanding bias in machine learning. *CoRR*, abs/1909.01866, 2019.

[16] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[17] Julio C. S. Jacques Junior, Yagmur Güçlütürk, Marc Pérez, Umut Güçlü, Carlos Andújar, Xavier Baró, Hugo Jair Escalante, Isabelle Guyon, Marcel A. J. van Gerven, Rob van Lier, and Sergio Escalera. First impressions: A survey on vision-based apparent personality trait analysis. *IEEE Transactions on Affective Computing (TAC)*, 2019.

[18] Julio C. S. Jacques Junior, Cagri Ozcinar, Marina Marjanovic, Xavier Baró, Gholamreza Anbarjafari, and Sergio Escalera. On the effect of age perception biases for real age regression. In *International Conference on Automatic Face & Gesture Recognition (FG)*, pages 1–8, 2019.

[19] Heinrich Jiang and Ofir Nachum. Identifying and correcting label bias in machine learning. *CoRR*, abs/1901.04966, 2019.

[20] J. Joo, F. F. Steen, and S. C. Zhu. Automated facial trait judgment and election outcome prediction: Social dimensions of

face. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3712–3720, 2015.

[21] Allison Langer, Ronit Feingold-Polak, Oliver Mueller, Philipp Kellmeyer, and Shelly Levy-Tzedek. Trust in socially assistive robots: Considerations for use in rehabilitation. *Neuroscience & Biobehavioral Reviews*, 104:231 – 239, 2019.

[22] L. Liang, L. Lin, L. Jin, D. Xie, and M. Li. SCUT-FBP5500: A diverse benchmark dataset for multi-paradigm facial beauty prediction. In *International Conference on Pattern Recognition (ICPR)*, pages 1598–1603, 2018.

[23] G. William Lucker, William E. Beane, and Robert L. Helmreich. The strength of the halo effect in physical attractiveness research. *The Journal of Psychology*, 107(1):69–75, 1981.

[24] S. C. Matz, M. Kosinski, G. Nave, and D. J. Stillwell. Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 114(48):12714–12719, 2017.

[25] DongWon Oh, Elinor A. Buck, and Alexander Todorov. Revealing hidden gender biases in competence impressions of faces. *Psychological Science*, 30(1):65–79, 2019.

[26] Carl L. Palmer and Rolfe D. Peterson. Halo effects and the attractiveness premium in perceptions of political expertise. *American Politics Research*, 44(2):353–382, 2016.

[27] VP Ponce-Lopez, B Chen, A Places, M Oliu, C Corneanu, X Baro, HJ Escalante, I Guyon, and S Escalera. ChaLearn LAP 2016: First round challenge on first impressions - dataset and results. In *European Conference on Computer Vision Workshop (ECCVW)*, pages 400–418, 2016.

[28] R. D. Pérez Principi, C. Palmero, Julio C. S. Jacques Junior, and S. Escalera. On the effect of observed subject biases in apparent personality analysis from audio-visual signals. *IEEE Transactions on Affective Computing*, pages 1–14, 2019.

[29] Novi Quadrianto, Viktoriia Sharmanska, and Oliver Thomas. Discovering fair representations in the data domain. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[30] Silvia Riva, Ezekiel Chinyio, and Paul Hampton. Biased perceptions and personality traits attribution: Cognitive aspects in future interventions for organizations. *Frontiers in Psychology*, 9, 2019.

[31] Joseph P. Robinson, Gennady Livitz, Yann Henon, Can Qin, Yun Fu, and Samson Timoner. Face recognition: Too bias, or not too bias? In *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020.

[32] Judy Hanwen Shen, Agata Lapedriza, and Rosalind W. Picard. Unintentional affective priming during labeling may bias labels. In *International Conference on Affective Computing and Intelligent Interaction*, 2019.

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015.

[34] Tomáš Sixta, Julio C. S. Jacques Junior, Pau Buch-Cardona, Eduard Vazquez, and Sergio Escalera. Fairface challenge at eccv 2020: Analyzing bias in face recognition. In *European Conference on Computer Vision Workshop (ECCVW)*, 2020.

[35] Sean N. Talamas, Kenneth I. Mavor, and David I. Perrett. Blinded by beauty: Attractiveness bias and accurate perceptions of academic performance. *PLOS ONE*, 11(2):1–18, 2016.

[36] Karl Timmerman and Jay Hewitt. Examining the halo effect of physical attractiveness. *Perceptual and Motor Skills*, 51(2):607–612, 1980.

[37] A. Todorov. *Face Value: The Irresistible Influence of First Impressions*. Princeton and Oxford: Princeton University Press, 2017.

[38] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1521–1528, 2011.

[39] Alessandro Vinciarelli and Gelareh Mohammadi. A survey of personality computing. *IEEE Transactions on Affective Computing (TAC)*, 5(3):273–291, 2014.

[40] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[41] Shen Yan, Di Huang, and Mohammad Soleymani. Mitigating biases in multimodal personality assessment. In *International Conference on Multimodal Interaction (ICMI)*, page 361–369, 2020.

[42] Seyma Yucer, Samet Akcay, Noura Al-Moubayed, and Toby P. Breckon. Exploring racial bias within face recognition via per-subject adversarially-enabled data augmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020.

[43] Leslie Zebrowitz and Robert Franklin. The attractiveness halo effect and the babyface stereotype in older and younger adults: Similarities, own-age accentuation, and older adult positivity effects. *Experimental aging research*, 40:375–393, 2014.

[44] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[45] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *CoRR*, abs/1707.09457, 2017.