The Seventh AAAI Conference on Human
Computation and Crowdsourcing (HCOMP-19)

How Do We Talk about Other People? Group (Un)Fairness in Natural Language Image Descriptions

Jahna Otterbacher,[1,2] Pınar Barlas,[2] Styliani Kleanthous,[1,2] Kyriakos Kyriakou[2]
[1]Open University of Cyprus
[2]Research Centre on Interactive Media, Smart Systems and Emerging Technologies
Nicosia, CYPRUS
{j.otterbacher, p.barlas, s.kleanthous, k.kyriakou}@rise.org.cy

Abstract

Crowdsourcing plays a key role in developing algorithms for image recognition or captioning. Major datasets, such as MS COCO or Flickr30K, have been built by eliciting natural language descriptions of images from workers. Yet such elicitation tasks are susceptible to human biases, including stereotyping people depicted in images. Given the growing concerns surrounding discrimination in algorithms, as well as in the data used to train them, it is necessary to take a critical look at this practice. We conduct experiments at Figure Eight using a controlled set of people images. Men and women of various races are positioned in the same manner, wearing a grey t-shirt. We prompt workers for 10 descriptive labels, and consider them using the human-centric approach, which assumes reporting bias. We find that "what's worth saying" about these uniform images often differs as a function of the gender and race of the depicted person, violating the notion of group fairness. Although this diversity in natural language people descriptions is expected and often beneficial, it could result in automated disparate impact if not managed properly.

Labels are for clothing. Labels are not for people.
–Martina Navratilova

Introduction

The emergence of the field of Fairness, Accountability and Transparency (FAT*) has led to an appreciation of machine learning and algorithmic systems as being socio-technical in nature (Barocas, Hardt, and Narayanan 2018). Every step of the development and evaluation process involves human judgment. Training and evaluation datasets attempt to capture some aspects of the state-of-the-world, and learning mechanisms are applied to create a model, often for predictive purposes. However, given the diversity of the world, its complexity and messiness, it is unsurprising that algorithmic systems reflect the biases prevalent in the societies in which they are trained, evaluated and deployed.

Computer vision provides many examples of the challenges in developing systems that treat people *fairly*. Given recent advances, algorithms for visual recognition are now commonplace in our information ecosystem. In dating apps, they are used to track aesthetic preferences, to serve as "visual matchmakers."[1] Other applications operate in domains where results directly affect lives; for example, in autonomous vehicles,[2] or in fighting child trafficking.[3]

However, there is growing documentation of socially biased behaviors in image analysis algorithms. One recent study found an increased error rate in gender classification for people with darker skin (as compared to lighter skin) and women (as compared to men), where the disparity in error rates exceeded 30% (Buolamwini and Gebru 2018). Another found that Black men were more likely to be tagged with a negative emotion than White men, when using Face++ and Microsoft's Face API (Rhue 2018). Finally, in a study of commercial image tagging services (Kyriakou et al. 2019), images of Black people were less likely to be described as being attractive, as compared to Whites or Asians.

In short, image recognition algorithms do not always treat people *fairly*. Given the increasing influence of this technology in our lives, it is critical to minimize its unwanted social biases. In order to do so, training datasets built via crowdworking platforms must be analyzed to better understand the nature of these biases.

Social bias in training data

Many biases observed in system output can be traced back to the training data. Language processing researchers were among the first to scrutinize visual recognition datasets, in which descriptions are expressed through text. van Miltenburg disputed the idea that crowdsourced descriptions would be based only on image content (van Miltenburg 2016). He quotes the researchers who built Flickr8K, who claim that by asking workers to describe the content of an image, without any information about its context, "we were able to obtain conceptual descriptions that focus only on the information that can be obtained from the image alone," (Hodosh, Young, and Hockenmaier 2013) (p. 859).

[1] https://blog.clarifai.com/4-ways-ai-is-improving-dating apps

[2] https://blog.clarifai.com/clarifai-featured-hack-val.ai-is-a-parking-app-for-your-self-driving-car

[3] https://aws.amazon.com/blogs/machine-learning/thorn-partners-with-amazon-rekognition-to-help-fight-child-sexual-abuse-and-trafficking/

Through an analysis of Flickr30K, van Miltenburg subsequently showed that workers make various inferences on images depicting people, which do not logically follow from the content of the image. In particular, he noted cases of gender, racial and ethnic stereotypes, as well as other "unwanted inferences" (e.g., ethnicity marking, suggesting that images of White people are the default) (van Miltenburg 2016).

Others have cited similar concerns with MS COCO, also generated by asking MTurk workers to provide a caption that "describe[s] all the important parts of the scene." (Chen et al. 2015). Zhao and colleagues documented rampant gender biases in the data for multilabel object classification as well as semantic role labeling (Zhao et al. 2017). For instance, images of women were often associated with labels surrounding activities such as cooking or shopping, while men were depicted as playing golf or driving. Hendricks and colleagues addressed similar issues but with respect to the image captioning task in MS COCO (Hendricks et al. 2018).

Berg and colleagues emphasized that crowd annotations are "human-centric annotations," which provide a wealth of information about how people judge images, and what they find important (Berg et al. 2012). In facing the problem of biased image data, Misra and colleagues adopted this approach (Misra et al. 2016). They explained that even the act of taking a photograph in our visually rich world exhibits a human reporting bias; we choose what is interesting enough to capture. Similarly, crowdworkers' descriptive labels on an image tell us "what's worth saying," but cannot be interpreted as a faithful description of image content.

Human-centric annotations: are they fair?

Even if we accept that biases are unavoidable in human-centric annotation, this does not mean that the resulting data is *fair*. With few exceptions (e.g., (Otterbacher 2018)), previous work has considered social biases in datasets of images collected in the wild (e.g., social media), in a variety of contexts. Thus, it is difficult to understand what triggers the biased descriptions. In contrast, our work examines how workers describe, in their own words, a set of highly-controlled images from the Chicago Face Database.[4] Since the images are devoid of context, the only characteristics that might trigger a biased response relate to the physical appearance and demographic attributes (race, gender) of the person.

As shown in Figure 1, which provides example images, along with crowd-generated tags (in Table 1), we observe that people images can be described in infinite ways. Some workers provide very concrete labels (e.g., face, eyes, neck, white background). Others, even if asked to describe image "content," make inferences; we observe abstract traits (e.g., happy, calm, diligent), characterizations of a person's demographic attributes (e.g., African American, woman, young, transgender) that cannot be confirmed, and even subjective judgments of the person (e.g., beautiful, nice, normal).

Given the wide variety of tags used, we propose not to examine specific words, but instead, to gauge workers' overall approach to the task. We consider the extent to which they use tags such as the above, which do not follow directly from the image content, and the extent to which workers' approach is correlated to the demographic attributes of the depicted person. We frame our analysis in terms of *group fairness*, addressing the following research questions:

- Q1. What's "worth reporting" in an image, in terms of sensitive attributes, and is group fairness respected?
- Q2. If sensitive or abstract tags are used, do they appear early on or later on in workers' responses?
- Q3. Are abstract tags more likely to be used when workers describe an in-group member?

Related Work

Datasets for developing visual recognition systems exhibit a range of biases. We review related literature to understand why social biases might occur, even when workers describe highly uniform images of faces, without context. We will also describe the notion of *fairness* in machine learning.

Interpreting human faces

People make inferences about others, even without contact or context. When shown images of strangers, people infer abstract characteristics, such as traits, automatically and almost immediately (Willis and Todorov 2006). Furthermore, abstract characteristics are interrelated to one's interpretation of global characteristics of faces (e.g., babyfacedness). Using photographs of people with neutral expressions, researchers demonstrated that physiognomic information (i.e., participants' inferences about traits) could change their perceptions of individuals, despite physical evidence to the contrary (Hassin and Trope 2000). Thus, we can expect crowdworkers to make inferences on images of people's faces; the question is if and how they express them.

Fiske and Cox studied people's descriptions of others, to determine which types of "person concepts" are used when describing strangers versus friends (Fiske and Cox 1979). Resonating with the notion of human-centric annotation, they noted a preference for open, textual responses, in contrast to closed responses, which have "limited our knowledge of the descriptive process." They found that use of physical attributes occurred more frequently when participants described a stranger, while familiar targets' descriptions were more interpretive. However, they noted differential usage of concepts during the course of creating a description, with concrete concepts appearing earlier on as compared to abstract. With specific respect to image tagging for personalization, it has been found that the order in which tags are provided is indicative of importance to the user (Nwana and Chen 2016). Therefore, we shall consider not only the types of tags workers provide when describing people images, but also the order in which they do so.

Bias in natural language descriptions

Rudinger and colleagues argued that tasks in which image descriptions are elicited through natural language, are particularly susceptible to socio-cognitive biases (Rudinger, May, and Van Durme 2017). They examined the Stanford Natural

[4] https://chicagofaces.org/default/

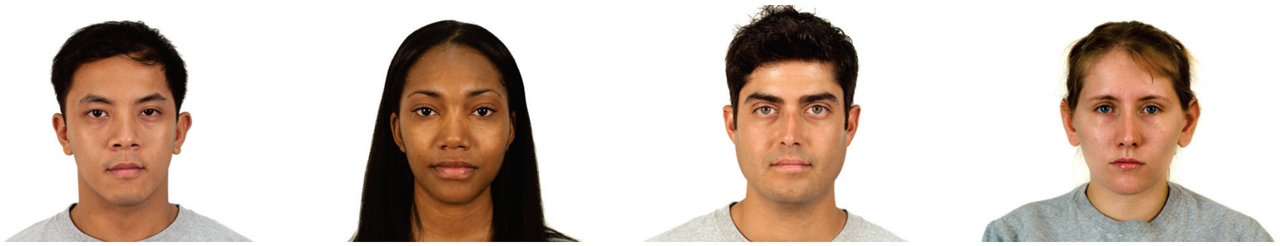

Figure 1: Four images from the Chicago Face Database (CFD) (left to right: AM-253, BF-233, LM-220, WF-036).

	India (IN)	United States (US)
AM-253	eyebrows, beautiful, yellow skin, shaved beard, hair, lips, grey tshirt, asian, black eyes, dark eyes, young, crew neck, long ears, short hair, eyes, dark hair, thick eyebrows, sexy, brunet, handsome, tshirt, face, nose, man, ears	good skin, short hair, brown hair, calm, nice looking, black hair, grey sweatshirt, young, grey shirt, oval face, non-american, asian man, clean shaven, big ears, male, straight brows, slight mustache, tshirt, brown eyes
BF-233	beautiful, black, black eyes, caring, diligent, eyes, girl, girl forehead, happy, lip, long hair, mouth, neck, nose, straight hair, wide nose, woman	black, brown eyes, chin, dark eyes, dark skin, ears, eyebrows, eyes, hair, lips, long hair, long neck, nice, normal, nose, nostrils, serious expression, shirt, straight hair, thin eyebrows
LM-220	eyebrows, medium ears, grey tshirt, brown skin, black hair, pink lips, eyes, hair, small mouth, chin, grey eyes, head, wheatish skin, fat, grey shirt, thick eyebrows, man, neck, thick lips, face, sharp nose, normal personality, small eyes, nose, ears	cute, black hair, big nose, thick eyebrows, wide jaw, abundant hair, big ears, tan, shirt, light eyes, young, long strands, spike hair, small front, short hair, long hair, brown, dark hair, green eyes, man, wide ears, plain, grey, thin, serious, male, round face
WF-036	beauty, big ears, blond, blue eyes, boy, brown hair, fair, female, good looking, long ears, shirt, short neck, straight hair, straight nose, thin lips, white, wide eyes, woman	blue eyes, brown hair, caucasian, front view, girl, lip gloss, long hair, plain expression, round face, short bangs, sober, solo, white background, woman, young

Table 1: Tags for CFD images provided by Figure Eight workers in two regions.

Language Inference corpus. In building the corpus, workers were presented with a caption from a Flickr image, and asked to generate additional captions. The Rudinger analysis showed that captions exhibit ethnic, racial and gender stereotypes. They noted that researchers should be very cautious of the elicitation protocols used.

Linguistic biases can be observed in tasks where participants describe images of people. Linguistic Expectancy Bias (LEB) predicts that we describe counter-stereotypical people more concretely as compared to more stereotypical people (Beukeboom et al. 2014). Stereotypical people, whose appearance is more expected, tend to be described abstractly, with inferences made by the annotator. This is also true of in-group members, who are more familiar. Evidence of LEB was documented in the ESP Game dataset (Otterbacher 2015), and in Flickr30K (van Miltenburg 2016).

The above findings resonate with previous work suggesting that generally, natural language data (e.g., texts collected from the Web) are subject to reporting bias. Co-occurrences reported in text often do not respond to their frequency in the offline world, as people may over- or under-report something unusual or counter-stereotypical (e.g., a "male nurse") (Bolukbasi et al. 2016). Gordon and Van Durme offer the example of learning about "human body parts" from online text (Gordon and Van Durme 2013). Although healthy people generally have a head and eyes, as well as a pancreas and a gallbladder, the latter two parts are rarely mentioned in text as compared to the former two parts.

Assessing fairness

The machine learning community has considered ways to mitigate bias in data and algorithms. There are two main approaches: *discrimination discovery* from datasets, and the prevention of discrimination using *fairness-aware learning* methods (Hajian, Bonchi, and Castillo 2016). Our work is inspired by the first approach, as we seek to understand if the race and gender of a person depicted in an image may influence annotators' approach to describing it.

Fairness is difficult to define; it can be understood as a placeholder term that relates to normative egalitarian considerations (Binns 2018). Nonetheless, two main notions have emerged in the literature: group and individual fairness. *Group fairness* holds that advantaged and protected groups should be treated in a similar manner (Pedreschi, Ruggieri, and Turini 2009). In classification or ranking tasks, group fairness can be understood as statistical parity; a minority group should receive the same treatment as the majority and/or the whole population. *Individual fairness* requires that similar individuals be treated in a consistent manner.

Our study focuses on group fairness. We assume that crowdworkers should take the same general approach when describing all images of people. Specifically, under the human-centric annotations approach, we shall see if "what's worth saying" about people images tends to be the same across eight social groups (two genders * four races). However, given the open-ended nature of the task, evaluating group fairness in image tagging is not as straight-forward as in a task (e.g., classification) with a closed set of outcomes. Therefore, in the next section, we develop a methodology through which we consider the *themes* mentioned when describing an image of a person.

Methodology

In this section, we provide details on the target images and the annotation task conducted via Figure Eight.[5] Following that, we detail the thematic coding of worker responses. Finally, we provide a summary of the general approach workers take to the open-ended task, before moving onto the question of whether their responses are fair.

Chicago Face Data

The Chicago Face Database (Ma, Correll, and Wittenbrink 2015) contains "high-resolution, standardized photographs of male and female faces of varying ethnicity" between the ages of 17-40. Each model is wearing the same t-shirt, standing in front of a white background, looking straight at the camera. We used the 597 portraits with neutral expressions. The distribution of the depicted persons' gender and race, self-reported from two and four mutually-exclusive categories respectively, is detailed in Table 2.

	Asian	**Black**	**Latino/a**	**White**	**Total**
Women	57	104	56	90	307
Men	52	93	52	93	290
Total	109	197	108	183	597

Table 2: Number of images by person's race and gender.

Crowd annotation task

We set up two tasks on Figure Eight, one each targeting workers in India and the U.S. These regions were chosen as they are among the largest, Anglophone pools of workers on the platform (Posch et al. 2018). The instructions asked workers to "help us determine the content in the images," by providing "individual words or two-word phrases" that "best describe the image content." Workers were first asked a simple question that served as an attention check ("Are there any humans in the image?") They were then asked to provide 10 words or phrases to describe the image. Finally, they were asked to provide their own gender and race, where the choices were those used in the CFD, as well as "other."

We collected three responses per image from unique workers; a worker could describe up to 20 images. Workers in India were paid 20 cents per image, while workers in the U.S. received 30 cents per image. As the task took no longer than 120 seconds, this corresponds to an hourly wage of 6 and 9 USD, respectively. Workers were satisfied with the job, both in terms of the set-up as well as the pay; in the Contributor Satisfaction survey, our India task received a rating of 4.7 out of 5 (n=27 respondents), while the U.S. task was rated 4.9 out of 5 (n=28 respondents).

	India	**U.S.**
Unique workers	107	116
Median time on task	120 seconds	120 seconds
Maximum time on task	27 minutes	29 minutes

Table 3: Summary statistics for crowdwork.

We could not use test questions for quality control, due to the open-ended nature of the task, but we did enforce a minimum time per image of 40 seconds. In addition, we used validators, regular expressions to ensure that one-to-two words were provided. However, we still had some "Fast Deceivers" (i.e., gibberish responses) (Gadiraju et al. 2015). 88 (5%) of the responses from the India workers' observations were resubmitted for work. The final dataset is presented in Table 4.

Worker region:	**India**		**U.S.**	
Image demographics:	M	W	M	W
Asian	156	167	156	171
Black	276	312	279	312
Latino/a	156	168	156	168
White	279	270	279	270

Table 4: Annotated images by workers in each region.

Thematic coding of tags

Because we want to examine "what's worth mentioning," we compare the themes used, rather than particular words. In (Barlas et al. 2019), we described an evaluation of all the tags collected. 21 themes were discovered, grouped into five major clusters. Currently, we consider three clusters: Demographics, Concrete and Abstract. We also consider some sub-clusters, which are pertinent to the fair treatment of people images. The clusters considered are described in Table 5. Names have been simplified for convenience; however, clusters often refer to concepts broader than the title (e.g.,"Race" includes adjacent concepts such as nationality and religion). Furthermore, clusters are not mutually-exclusive; if a tag ("lady") implies more than one concept (feminine and adult age), it appears in all applicable clusters. Workers' descriptions were coded using this typology. Our dataset, including the typology and dictionaries, is publicly available.[6]

General approach of workers

Table 6 examines the approach of crowdworkers in terms of the types of attributes "worth mentioning" across all CFD images. Here, we consider two sets of tags: all 10 tags versus the very first tag provided. In the full set of tags, crowdworkers from each region tend to cover all three thematic clusters,

[5] https://www.figure-eight.com/

[6] https://doi.org/10.7910/DVN/APZKSS

Cluster	Description	Examples
Demographics	*Tags that describe the inferred gender, age and/or origin(s) of the depicted person*	
Gender	Tags that refer to a gender identity or expression	masculinity, girl, androgynous
Age	Tags that refer to the person's age	millenial, girl, thirties
Race	Tags that refer to the person's race, ethnicity, nationality, or religion	nigerian, migrant, Muslim
Concrete	*Tags that describe directly observable attributes of the image or the depicted person*	
Shape	Tags that refer to the shape, size/amount, or position of the person or something about the person	crooked_nose, fat, nonsymmetrical
Abstract	*Tags that describe the inferred, subjective or conceptual attributes of the person*	
Judgment	Tags that describe an opinion or subjective description	normal, beautiful, photogenic
Traits	Tags that refer to a personality trait or enduring characteristic	extrovert, stubborn, macho
Emotion	Tags that refer to an emotional, mental, or temporary physical state	happy, concentrated, serious

Table 5: Thematic cluster names, explanations, & example tags used in the present study.

	All tags		First tag	
	IN	US	IN	US
Demographic	0.80	0.83	0.59	0.64
Gender	0.68	0.74	0.53	0.56
Age	0.62	0.63	0.41	0.42
Race/ethnicity	0.38	0.41	0.06	0.07
Concrete	0.93	0.99	0.30	0.35
Shape	0.67	0.75	0.09	0.06
Abstract	0.51	0.55	0.05	0.03
Judgment	0.26	0.22	0.03	0.01
Traits	0.21	0.18	0.01	0.007
Emotion	0.26	0.36	0.01	0.008

Table 6: Proportion of descriptions using at least one tag from the given cluster: all 10 (left), first tag only (right).

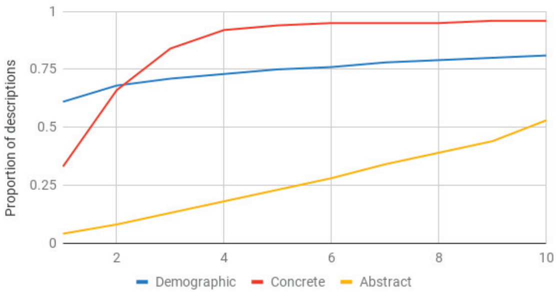

Figure 2: Prop. of descriptions (from 1st to all 10 tags) including one or more tags from each cluster.

and almost always mention Concrete (i.e., directly observable) attributes. In over 80% of responses, they have inferred at least one Demographic attribute, whereas in half of the descriptions they also include Abstract inferences. Furthermore, most annotations involve a mix of attributes.

Around 60% of the first tags workers provide refer to Demographic attributes of the depicted individual, typically gender and/or age, while around 30% are Concrete. Figure 2 examines the use of tags from the three major clusters. The proportion of descriptions using at least one tag from the cluster is plotted, when we consider only the first tag provided by the worker, up until the full set of 10 tags. We observe that Demographic tags are typically mentioned in the first couple of tags. Similarly, Concrete tags are provided up front, and in the first five tags, nearly all descriptions contain at least one Concrete tag. The use of Abstract tags increases the more tags we consider, however, over all 10 tags, only half of the descriptions use them.

Analysis

We now look deeper into our data to see if and how "what's worth saying" differs across social groups. We consider seven types of tags: Gender, Age, Race, Shape, Judgment, Traits and Emotion. These tags appear frequently in the descriptions, as shown in Table 6; however, they indicate a degree of inference on the part of the worker, as compared to concrete attributes (e.g., clothing or colors). "Shape," although a concrete attribute in our schema, when applied to CFD images, indicates particular detail to a person's appearance (e.g., "crooked nose," "fat," "nonsymmetrical").

Q1. "What's worth saying" about people images?

Model. To assess "what's worth saying" and if workers' descriptions respect group fairness, we compare the likelihood of the seven types of tags being used, when describing an image of a White man (WM), as compared to the other seven social groups (White woman (WW), Asian woman/man (AW/AM), Latina/o (LW/LM) Black woman/man (BW/BM)). For each type of tag, we fit a logistic regression model, which predicts the event that at least one such tag appears in a worker's description, as a function of the race and gender of the depicted person. We include main effects for these demographic attributes, along with their interaction. We also include a main effect on worker region (IN/US). Because we have at least five descriptions for each image, we include random effects, using R's lme4 package for linear mixed-effects models.[7]

In the event that a logistic regression model reveals that

[7] https://cran.r-project.org/web/packages/lme4/lme4.pdf

	Intercept/WM	AM	AW	BM	BW	LM	LW	WW	Region/US
Gender	0.941***	-0.197	-0.593*** (0.55)	-0.260	-0.3196* (0.73)	0.0665	0.0304	-0.111	0.296*** (1.34)
Age	0.555***	0.206	-0.562$* * *$ (0.57)	-0.108	-0.164	0.162	0.132	-0.0517	0.0389
Race	-0.764***	0.886*** (2.43)	1.179** (3.25)	0.397** (1.49)	0.175	-0.121	0.0239	-0.237	0.150* (1.16)
Shape	0.608***	-0.016	0.653*** (2.26)	0.188	0.274	-0.078	0.235	-0.153	0.385*** (1.47)
Judgment	-1.157***	-0.0919	0.282	-0.328* (0.72)	0.250	0.165	0.407* (1.50)	0.346* (1.41)	-0.230** (0.80)
Traits	-1.207***	-0.171	-0.466	-0.148	-0.225	0.080	0.0712	-0.149	-0.157
Emotion	-1.098***	0.118	-0.260	0.143	0.140	-0.004	0.217	0.0740	0.433*** (1.54)

Table 7: Logistic regression models to predict use of one or more tags describing sensitive attributes.

	WM	AM	AW	BM	BW	LM	LW	WW	Sig. diff. (over all data per Tukey)
Gender-IN	0.74	0.67	0.60	0.67	0.62	0.71	0.73	0.70	AW < LM,LW,WW,WM
Gender-US	0.75	0.74	0.64	0.72	0.75	0.81	0.77	0.75	
Age-IN	0.69	0.68	0.47	0.61	0.56	0.67	0.66	0.69	AW < AM,BM,LM,LW,WM,WW
Age-US	0.59	0.69	0.54	0.62	0.64	0.68	0.68	0.59	
Race-IN	0.29	0.51	0.62	0.44	0.35	0.31	0.29	0.30	AM,AW > WM,WW,BM,BW,LM,LW
Race-US	0.39	0.58	0.62	0.42	0.40	0.31	0.39	0.28	BM>WM,WW,LM
Shape-IN	0.64	0.63	0.81	0.72	0.68	0.58	0.70	0.60	AW > AM,LM,WM,WW
Shape-US	0.73	0.73	0.79	0.72	0.79	0.75	0.76	0.70	
Judgment-IN	0.23	0.21	0.32	0.18	0.28	0.30	0.35	0.30	AW > BM
Judgment-US	0.21	0.21	0.23	0.16	0.25	0.20	0.25	0.27	BM < BW,LW,WW
Traits-IN	0.24	0.24	0.11	0.13	0.20	0.29	0.22	0.26	
Traits-US	0.19	0.13	0.18	0.25	0.16	0.17	0.16	0.20	
Emotion-IN	0.29	0.29	0.21	0.24	0.25	0.28	0.32	0.26	
Emotion-US	0.30	0.35	0.28	0.41	0.39	0.32	0.37	0.37	

Table 8: Proportion of descriptions, by social group, using at least one sensitive or abstract tag & results of post-hoc test.

not all groups are treated in the same manner as White men, we conduct a post-hoc test. Specifically, pairwise comparisons are made using a Tukey test, implemented in R's multcomp.[8] We report any differences with a p-value < 0.05.[9]

Table 7 details the model for predicting the use of each of the seven types of tags, based on the set of 10 tags provided by workers. Odds ratios are provided as a measure of effect size for coefficients that are statistically significant (i.e., the use of a tag differs for a given social group, as compared to WM). Table 8 details, for each group, the proportion of descriptions that include at least one tag of a given type. In addition, the far-right column details the pairwise comparisons that are significant per the post-hoc Tukey test.

Observations. From Table 7, we observe that the main effect on worker region is significant for some tags. US workers tend to make greater use of Gender, Race, Shape and Emotion tags, whereas workers in India use more Judgment. Therefore, region was left in the model as a control variable. From Table 8, we observe significant differences between some social groups; "what's worth saying" about images of Asians and Blacks often differs from the other groups.

With respect to images of AW, workers are less likely to mention gender and age when describing them, as compared to four other groups (LM, LW, WW, WM). However, along with images of AM, these images receive more tags that describe the individual's race or ethnicity, as compared to all other social groups. Finally, Asian women receive more tags that describe shape (e.g., "thin eyebrows," "round face").

Images of Black individuals are more often marked with race-related tags, as compared to Whites and Latino men. Interestingly, Black men receive judgment-related tags less often than other groups; this difference is statistically significant when compared to Black, Latina and White women.

In summary, the attributes mentioned are not independent of the race and gender of the person being described. Our results are in line with previous observations that workers view White individuals as the default, with others receiving marked descriptions (van Miltenburg 2016). Although an individual's gender is noted in more than half of the first tags that workers use (as noted in Table 6), it appears that workers find other attributes more "worth noting" in the case of images of AW, such as the shape of one's facial features.

[8] cran.r-project.org/web/packages/multcomp/multcomp.pdf

[9] We use the following conventions for reporting statistical significance: *** p<.001, ** p<.01, * p<.05.

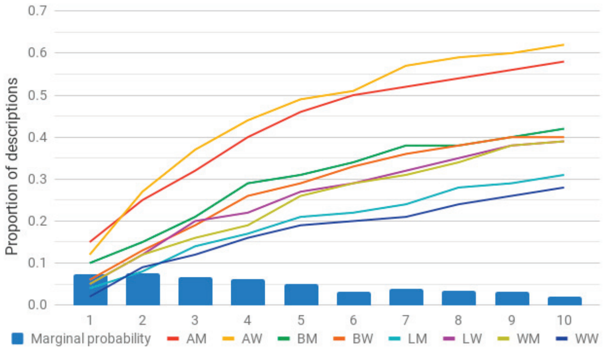

Figure 3: Prop. of US descriptions including a Race tag.

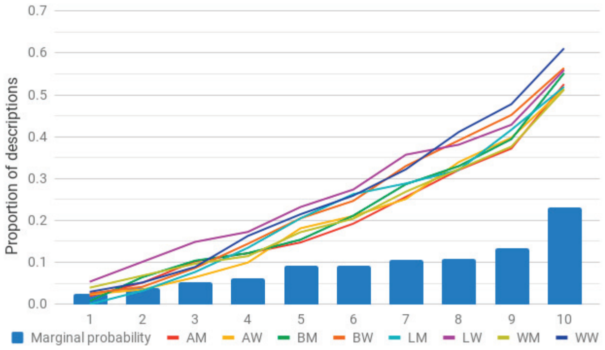

Figure 4: Prop. of US descriptions including an Abstract tag.

Q2. When are sensitive and abstract tags used?

Next, we seek to understand when in the annotation process workers introduce sensitive tags. Having seen that US and India-based workers show different behaviors with respect to the use of tags describing sensitive and abstract characteristics, we now focus on the US workers. Given that the CFD images were collected in the US context, we chose to focus on the US workers and their use of Race and Abstract tags, which includes Judgment, Traits and Emotion.

Race. Figure 3 plots the proportion of descriptions that include at least one Race tag during the course of the task (i.e., from Tag 1 to Tag 10). The marginal distribution of Race tags at each point in time is shown over all images, as well as the cumulative distribution over time, broken out by the eight social groups. Considering the marginal probabilities, we observe that Race tags are usually provided early on in the task, and are used less frequently as time goes on.

Considering the cumulative distribution, it is clear that Asian images receive more Race tags at the beginning of the task (15% for AM, 12% for AW at Tag 1). In comparison, for WM/WW the rates are 5% and 2%, respectively. Although the proportion of descriptions with a Race tag increases over time for all groups, the trend is particularly pronounced for Asians. For instance, by the addition of the sixth tag, over half of the descriptions of Asians mentions race; in contrast, even over all 10 tags, no other group reaches 50%.

	US-WM	US-WW	z
AM	0.69	0.54	1.00
AW	0.56	0.52	0.29
BM	0.64	0.59	0.32
BW	0.63	0.57	0.55
LM	0.54	0.57	-0.31
LW	0.80	0.56	1.47
WM	0.48	0.57	-1.41
WW	0.82	0.61	1.94*

Table 9: Prop. of descriptions with an Abstract tag.

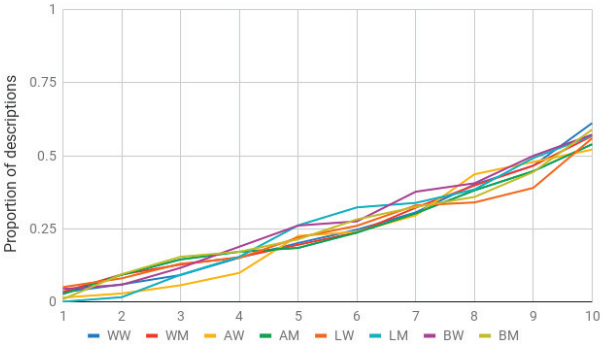

Figure 5: Descriptions by US-WW with an Abstract tag.

Abstract. Figure 4 details the proportion of descriptions using at least one Abstract tag to describe the target person. In comparison to Race tags, Abstract tags are added later on in the task. Only by the addition of the 10th tag, do we find that 50% of the descriptions contain an Abstract tag; this happens across all social groups. Descriptions of WW are the most likely to contain Abstract tags, however, this trend is only seen after the 8th tags are added to the descriptions.

Q3. Are in-group members described abstractly?

Finally, we analyze behaviors of the two largest groups of US workers, WW (n=904 images labelled) and WM (n=286). We consider the use of Abstract tags. Theory predicts that familiar individuals will be described more abstractly; therefore, we compare how WW and WM label their in-group members versus others. Table 9 compares the overall use of Abstract tags between WW and WM workers, using a Z-test for two population proportions. As shown, WM describe WW more abstractly than WW themselves.

White women workers. Figure 5 shows the use of Abstract tags by WW (cumulative distribution), over the course of the task, from Tag 1 to 10. We observe a nearly linear trend for all eight social groups. Although the in-group, images of WW, has the largest proportion of descriptions including an Abstract tag (61%), there is no clear tendency for this group to deviate from the others.

White men workers. Figure 6, which details WM's use of Abstract tags, shows a deviation for two groups of target individuals, WW and LW. In particular, by the 3rd tag, there

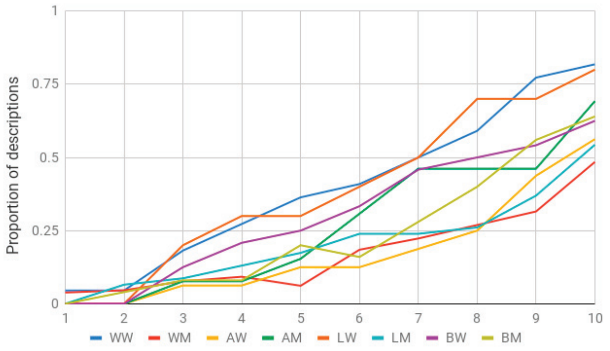

Figure 6: Descriptions by US-WM with an Abstract tag.

is a tendency for WM to use more Abstract tags for these groups. By the 10th tag, 82% of descriptions on WW, and 80% on LW, contain an Abstract tag. Thus, rather than an in-group/out-group effect, this suggests a cross-gender effect, with WM being more inferential when describing WW.

Discussion

Human-centric annotations on people images contain a wealth of information beyond the image content. This is expected; since people perceive others differently, based on their unique characteristics and social relation to the target person, they may describe them using a range of attributes, particularly when asked to do so through natural language. However, this diversity in "what's worth saying" about a set of images can become an issue when it reflects prevalent social biases (e.g., ethnicity marking, which positions Whites as the norm). If algorithms are trained on datasets such as ours uncritically, it could lead to automated unfairness.

In our study, crowdworkers were given the task of describing the content of highly uniform people images and we interpreted their responses as being human-centric. Thus, they tell us "what's worth mentioning" about each image. Despite the highly uniform nature of the images, and the lack of context, workers used Demographic and Abstract person attributes in various ways. While differences in usage of Concrete tags such as body parts (head, face) or colors (grey t-shirt, white background) may not be so important, differences in reporting sensitive attributes such as demographic and abstract characteristics, or facial shapes deviating from "the norm" can lead to harmful biases in automated systems.

It is extremely doubtful that annotations on people images would ever respect group fairness, without more structuring of the task and incentives. Our analysis confirmed what was suggested by the literature review: our open-ended task, in which workers were asked to provide natural language responses, exhibited social biases. In particular, Asian men and women, as well as Black men, were described differently as compared to other social groups. This is likely a population-wide bias (Kamar, Kapoor, and Horvitz 2015); when workers used their own words to describe target individuals, non-whites are marked with a Race attribute.

Furthermore, we found that Race tags are typically mentioned up front, among the first descriptive tags offered by workers, suggesting that workers find this an important attribute to report. However, Abstract tags having to do with inferred traits or emotions of a depicted person, or the worker's judgment of the person, tended to be added by workers later on, during the last of the 10 tags. Thus, when creating datasets, tag order contains valuable information in terms of the salience of the trait and should be preserved.

Developers and researchers who need to crowdsource datasets for visual recognition algorithms, should consider carefully how they elicit descriptions from workers, when images depict people. Stressing the need to "describe the image content" in the task does not result in workers limiting their descriptions to attributes that follow logically from image content. Instead, they may wish to consider more specific prompts, particularly if they want annotations to respect group fairness. Descriptions of people images may start out very concrete (i.e., based on content) but become more abstract, the longer the description is required to be (the more tags that are required). Depending on the intended use for the dataset, posing a shorter task, and/or altering workers' incentives might be a means to reduce responses that are characterized by the above biases (Faltings et al. 2014). Another solution might be to specifically prompt workers to provide a desired number of tags of a given type (e.g., "Please tell us two concrete observations about the image, as well as two demographic characteristics of the depicted individual").

Limitations

Our analyses used the binary response variable, whether or not a sensitive/abstract attribute was used in a description. Due to space constraints, we have not presented analyses on the continuous variable (i.e., to what extent an attribute was used); the results were parallel. Another analysis for future work should consider the combinations of features used in a description, rather than just one at a time. Still another aspect to be examined is the role of physical attractiveness of the depicted person in eliciting abstract descriptions.

Conclusion

When asked to provide descriptive labels on images of people, workers do not do so in a way that is necessarily fair. It is human nature to make inferences about others, subconsciously and immediately, even when the only information available is a photo of a face (Willis and Todorov 2006). Still, human-centric descriptions, which go beyond what's concretely in a image, and provide information on "what's worth saying" about it, contain a wealth of information, not only about the target image but also about the worker. Therefore, when building datasets that contain people images, we must think carefully and deeply about what information we really need to capture, and how it should be represented.

Acknowledgments

This work has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreements No 810105 (CyCAT) and 739578 (RISE) complemented by the Government of the Republic of

Cyprus through the Directorate General for European Programmes, Coordination and Development.

References

Barlas, P.; Kyriakou, K.; Kleanthous, S.; and Otterbacher, J. 2019. Social b(eye)as: Human and machine descriptions of people images. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, 583–591.

Barocas, S.; Hardt, M.; and Narayanan, A. 2018. Fairness and machine learning. fairmlbook. org, 2018. *URL: http://www. fairmlbook. org*.

Berg, A. C.; Berg, T. L.; Daume, H.; Dodge, J.; Goyal, A.; Han, X.; Mensch, A.; Mitchell, M.; Sood, A.; Stratos, K.; et al. 2012. Understanding and predicting importance in images. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 3562–3569. IEEE.

Beukeboom, C.; Forgas, J.; Vincze, O.; and Laszlo, J. 2014. Mechanisms of linguistic bias: How words reflect and maintain stereotypic expectancies. *Social Cognition and Communication* 313–330.

Binns, R. 2018. Fairness in machine learning: Lessons from political philosophy. In *Conference on Fairness, Accountability and Transparency*, 149–159.

Bolukbasi, T.; Chang, K.-W.; Zou, J. Y.; Saligrama, V.; and Kalai, A. T. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, 4349–4357.

Buolamwini, J., and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, 77–91.

Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollár, P.; and Zitnick, C. L. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Faltings, B.; Jurca, R.; Pu, P.; and Tran, B. D. 2014. Incentives to counter bias in human computation. In *Second AAAI conference on human computation and crowdsourcing*.

Fiske, S. T., and Cox, M. G. 1979. Person concepts: The effect of target familiarity and descriptive purpose on the process of describing others 1. *Journal of Personality* 47(1):136–161.

Gadiraju, U.; Kawase, R.; Dietze, S.; and Demartini, G. 2015. Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 1631–1640. ACM.

Gordon, J., and Van Durme, B. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, 25–30. ACM.

Hajian, S.; Bonchi, F.; and Castillo, C. 2016. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2125–2126. ACM.

Hassin, R., and Trope, Y. 2000. Facing faces: studies on the cognitive aspects of physiognomy. *Journal of personality and social psychology* 78(5):837.

Hendricks, L. A.; Burns, K.; Saenko, K.; Darrell, T.; and Rohrbach, A. 2018. Women also snowboard: Overcoming bias in captioning models. In *European Conference on Computer Vision*, 793–811. Springer.

Hodosh, M.; Young, P.; and Hockenmaier, J. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research* 47:853–899.

Kamar, E.; Kapoor, A.; and Horvitz, E. 2015. Identifying and accounting for task-dependent bias in crowdsourcing. In *Third AAAI Conference on Human Computation and Crowdsourcing*.

Kyriakou, K.; Barlas, P.; Kleanthous, S.; and Otterbacher, J. 2019. Fairness in proprietary image tagging algorithms: A cross-platform audit on people images. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, 313–322.

Ma, D. S.; Correll, J.; and Wittenbrink, B. 2015. The chicago face database: A free stimulus set of faces and norming data. *Behavior research methods* 47(4):1122–1135.

Misra, I.; Lawrence Zitnick, C.; Mitchell, M.; and Girshick, R. 2016. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2930–2939.

Nwana, A. O., and Chen, T. 2016. Who ordered this?: Exploiting implicit user tag order preferences for personalized image tagging. In *2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 1–6. IEEE.

Otterbacher, J. 2015. Crowdsourcing stereotypes: Linguistic bias in metadata generated via gwap. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, 1955–1964. New York, NY, USA: ACM.

Otterbacher, J. 2018. Social cues, social biases: stereotypes in annotations on people images. In *Sixth AAAI Conference on Human Computation and Crowdsourcing*.

Pedreschi, D.; Ruggieri, S.; and Turini, F. 2009. Measuring discrimination in socially-sensitive decision records. In *Proceedings of the 2009 SIAM International Conference on Data Mining*, 581–592. SIAM.

Posch, L.; Bleier, A.; Flöck, F.; and Strohmaier, M. 2018. Characterizing the global crowd workforce: A cross-country comparison of crowdworker demographics. *arXiv preprint arXiv:1812.05948*.

Rhue, L. 2018. Racial influence on automated perceptions of emotions. *Available at SSRN 3281765*.

Rudinger, R.; May, C.; and Van Durme, B. 2017. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 74–79.

van Miltenburg, E. 2016. Stereotyping and bias in the flickr30k dataset. In *Proceedings of the Workshop on Multimodal Corpora: Computer Vision and Language Processing (MMC-2016)*, 1–4.

Willis, J., and Todorov, A. 2006. First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological science* 17(7):592–598.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.