Social Cues, Social Biases: Stereotypes in Annotations on People Images

Jahna Otterbacher
Faculty of Pure and Applied Sciences, Open University of Cyprus
Research Centre on Interactive Media Smart Systems and Emerging Technologies
Nicosia, CYPRUS

Abstract

Human computation is often subject to systematic biases. We consider the case of *linguistic biases* and their consequences for the words that crowdworkers use to describe people images in an annotation task. Social psychologists explain that when describing others, the subconscious perpetuation of stereotypes is inevitable, as we describe stereotype-congruent people and/or in-group members more abstractly than others. In an MTurk experiment we show evidence of these biases, which are exacerbated when an image's "popular tags" are displayed, a common feature used to provide social information to workers. Underscoring recent calls for a deeper examination of the role of training data quality in algorithmic biases, results suggest that it is rather easy to sway human judgment.

Introduction

It is increasingly easy and inexpensive to exploit human computation in solving business and research problems. Using commercial platforms such as MTurk or Crowdflower has become particularly convenient; one can design a human intelligence task (HIT), recruit a workforce to complete it, and train algorithms based on the collected data, "with a few clicks" [1]. However, a model is only as good as the data upon which it is trained, and ensuring the validity of workers' contributions is often non-trivial.

In recent years, researchers have addressed the issue of *bias* in crowdwork from a number of perspectives. Some have focused on understanding the demographic characteristics of crowdworkers and their correlation to the quality of work produced. Ross and colleagues' study of MTurk demonstrated a tendency toward a young, well-educated workforce, primarily concentrated in the U.S. and India (Ross et al. 2010). Worker personality and motivation for participation in crowdwork have also received attention. For instance, Kazai and colleagues (Kazai, Kamps, and Milic-Frayling 2012) found that Big Five characteristics, such as openness and conscientiousness, correlate to work quality, and that certain workers appear to be attracted to tasks with particular design characteristics. Furthermore, they found that workers who are motivated by financial benefit and/or personal fulfillment tend to provide quality answers in a labeling task (Kazai, Kamps, and Milic-Frayling 2013).

Beyond worker attributes and motivations, the design of the task itself can lead to biased output. In a study of the citizen science platform, Galaxy Zoo, which guides participants in answering questions about images of celestial objects, Kamar and colleagues (Kamar, Kapoor, and Horvitz 2015) reported that up to 23% of tasks resulted in responses for which the majority answer from the crowd differed from ground truth. Furthermore, they were able to identify certain properties of images that lead to biased answers. Likewise, intra-task effects can also bias workers' answers. In other words, previous tasks in a batch often influence the annotations provided in later tasks (Zhuang and Young 2015).

While there is significant noise in the answers provided by individual workers (Ipeirotis, Provost, and Wang 2010), assigning a HIT to a large number of workers – or exploiting "crowd wisdom" – is often used as a means to mitigate this noise (Kamar, Hacker, and Horvitz 2012). However, in the case of task-dependent biases, such strategies may not be valid, as the majority response may be incorrect or problematic. In other words, certain task characteristics might induce a population-wide bias (Kamar, Kapoor, and Horvitz 2015).

In the current work, we undertake an exploration of biases that manifest in a visual perception task involving the annotation of *people images*. As a task, image annotation has been described as subjective and/or open to interpretation, thus requiring careful consideration of the task design and incentive mechanisms (Faltings et al. 2014). In the particular case of people images, social stereotypes may be perpetuated in workers' descriptions of the images. These stereotypes would then be carried downstream into the algorithms that are trained on this data (Zhao et al. 2017).

Indeed, social psychology theory suggests that annotation tasks on people images will be subject to population-wide *linguistic biases*. Linguistic bias can be defined as a systematic asymmetry in the way that one uses language, as a function of the social group of the person(s) being described (Beukeboom et al. 2014). When asked to describe a person image in their own words, the theory predicts that workers will tend to inadvertently choose words that reveal stereotypes and prejudices toward particular social groups. Furthermore, as will be explained, these biases could be ex-

[1] https://www.crowdflower.com/machine-learning-with-crowdflower/

acerbated by the context and incentives provided through the design features of the HIT.

Image annotation and task design

Many design decisions go into the creation of an effective HIT. On the one hand, contributors must be given enough information to complete the task in a manner that provides the needed data. However, their responses can also be biased by the question/problem formulation. Designers must also determine how to effectively incentivize contributions, while at the same time ensuring quality responses (von Ahn and Dabbish 2004).

A key challenge is that the inter-dependencies and tensions between these design decisions are "intricate and little understood" (Law and von Ahn 2011) (p. 58). In our work, we focus on how the problem is posed and specifically, whether or not task information, in the form of social cues (i.e., how others responded to the problem), is provided. As will be detailed, we evaluate two versions of a simple annotation task, performed by individual contributors, whom we recruit and compensate via Mechanical Turk. We do not study the problem of incentives or task gamification.

Open annotation, in which contributors describe an image in their own words rather than using a controlled vocabulary, offers the potential to capture what "average people" perceive an image to be or to portray (Ames and Naaman 2007). While no universal guidelines for evaluating annotation quality have emerged, most information professionals speak of accuracy and value (Manzo et al. 2015; Park 2009). Human computation researchers aim for correctness (Law and von Ahn 2011), which is typically evaluated by between-user agreement (Law and von Ahn 2009). Subjectivity has been found to be an issue that threatens the value of the collected data; self-referential responses may not reflect an underlying ground truth, given the differences in perception of individual contributors (Law and von Ahn 2011) (p. 25).

Adding a social element to an annotation task (e.g., showing words that others used to describe the image) provides information to contributors that could guide their judgments in desirable ways. For instance, being shown "popular" or "recommended" tags might discourage users from suggesting words that are self-referential (Rader and Wash 2008). In games-with-a-purpose (GWAPs) for image annotation, displaying previously used tags has been used to encourage contributors to generate more creative words (von Ahn and Dabbish 2008). However, there are unique issues surrounding annotation tasks on people-related multimedia. As we shall see, adding social information to the task, may also serve as a trigger for linguistic biases in contributors' responses, as it provides cues about the intended purpose and audience of the image description.

Goals of the current work

Previously, we demonstrated, through secondary data analysis, the prevalence of linguistic biases in existing datasets of person images annotated by crowdworkers (Anonymous). Specifically, it was found that workers systematically used more subjective and evaluative language to describe images of women as compared to images of men, in a way that perpetuates prevalent gender stereotypes. However, that study did not examine how worker demographics are correlated to this process, nor could it control the context in which the images were described by workers or the content of the images themselves. In contrast, the present work undertakes a controlled experiment on MTurk, to explore both gender- and race-based linguistic biases, as well as to better understand the importance of the task design. The research questions we address are two-fold:

- Do we observe systematic differences in the words used to describe people images, as a function of the demographic characteristics (gender, race) of the worker and of the depicted person?
- Does the presence of social information concerning others' word choices affect this process?

In the next section, we provide the theoretical motivation for our study and develop specific hypotheses to be tested.

Background

Here we provide the theoretical background on *linguistic bias*, detailing two subtypes of biases: the Linguistic Expectancy Bias (LEB) and the Linguistic In-group Bias (LIB).

Linguistic biases and their detection

Two types of biases may affect the way that contributors describe images of people, the LEB and the LIB. Both can be detected based on the extent to which a person description (e.g., set of words provided for an image) uses abstract, subjective language (Semin, de Montes, and Valencia 2003). In a concrete description, one documents observed actions or details, without making inferences. An example would be a contributor describing what the depicted individual is wearing, and what she is doing in the image. In contrast, an abstract description is general and interpretive, such that it applies across events and scenarios. According to Semin and Fiedler (Semin and Fiedler 1991), the most abstract descriptions tend to use adjective predicates (as opposed to verbs describing specific actions), which detail characteristics of a person (e.g., "He is professional and clean-cut" versus "He served a drink.")

Abstract descriptions are more powerful than concrete ones, as they imply stability over time as well as generalizability. It has been confirmed that message recipients are impacted by abstraction in the language used, with more abstract descriptions being interpreted as enduring characteristics of the target person, in contrast to concrete descriptions, which are seen as being transient (Wigboldus, Semin, and Spears 2000). Such systematic differences in image metadata could lead to the development of algorithms that are less likely to retrieve or detect minorities. For instance, if images of engineers from minority social groups are labeled with concrete details, and those of white men are labeled more abstractly, including interpreting the person's occupation as being an engineer, this could lead to biases.

An important caveat is that linguistic biases occur most often when the communication has a clear purpose (Wigboldus, Semin, and Spears 2006). In particular, the LIB is known to depend on context – it does not occur in the absence of a communicative purpose, where one cannot define the perceived audience (Semin, de Montes, and Valencia 2003). We now turn to describing each of the two types of biases, in order to better understand why they occur.

LEB. The LEB reflects that it is easiest for us to process information that is prototypical (e.g., an image depicting a stereotype-confirming individual) (Winkielman et al. 2006). Laboratory studies have repeatedly demonstrated that when participants are asked to describe a scene depicting someone who violates their expectations, that they are likely to focus on particular details, providing tangible and concrete information (Wigboldus, Semin, and Spears 2006). On the other hand, when presented with scenes depicting stereotype-congruent (i.e., expected) individuals, participants are more likely to provide abstract details, such as their perceptions of the depicted person's disposition or character traits.

LIB. The LIB builds on the LEB; we expect our in-group members to have positive qualities and behaviors, while we may not hold such expectations for out-group members. The LIB predicts that we use language in such a way that it is difficult to disconfirm our pre-existing ideas (Maass et al. 1989). Therefore, we are more likely to describe the positive actions and attributes of in-group members with abstract language, whereas any negative traits and actions are more likely described concretely. The converse is predicted for descriptions of out-group members.

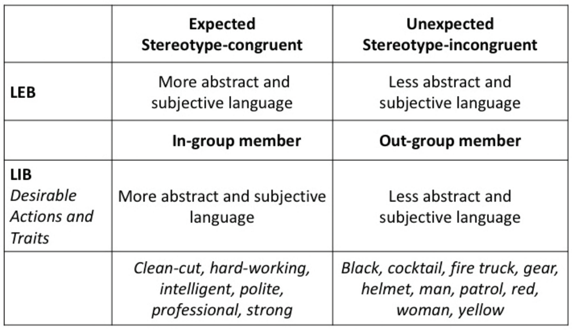
Detecting evidence of LEB and LIB. Figure 1 summarizes the features of image descriptions, which are predicted by the LEB and LIB. As can be seen, predictions are based on the social group of the depicted individual; in the case of the LEB, whether or not the individual is generally stereotype-congruent and in the case of the LIB, whether or not the individual is an in- or out-group member. Since all images we will test depict individuals in a positive, professional context, we do not evaluate the LIB in the case where the individual is shown in an undesirable manner. We also show example words used to describe our images of bartenders, firefighters and police officers. In the next section, we describe the manner in which we detect abstract and subjective language in image descriptions, as well as our experimental set-up.

Methodology

We conducted an experiment on MTurk, in which we manipulated the characteristics of the depicted persons in the images, as well as the *social information* provided to workers. We hypothesized that they would use more abstract language to describe stereotype-congruent images (i.e., LEB) and/or images of persons from their own social group (i.e., LIB), but that this effect would be mediated by the presence of social cues.

Images of target persons

We obtained a set of images of people in three working-class professions: bartenders, police officers and firefighters. For each profession, we used Shutterstock.com to find four images: a black woman, a black man, a white woman and a white man. All images had the following characteristics: (1) the person's face was clearly shown and he/she was smiling, (2) he/she was depicted in uniform, (3) there were no other people in the image. Candidate images were tested independently with five graduate students, who were asked to evaluate the above characteristics as well as the depicted person's race and gender, until we identified the final set of 12 images upon which there was full consensus.



	Expected **Stereotype-congruent**	**Unexpected** **Stereotype-incongruent**
LEB	More abstract and subjective language	Less abstract and subjective language
	In-group member	**Out-group member**
LIB *Desirable Actions and Traits*	More abstract and subjective language	Less abstract and subjective language
	Clean-cut, hard-working, intelligent, polite, professional, strong	*Black, cocktail, fire truck, gear, helmet, man, patrol, red, woman, yellow*

Figure 1: Person description characteristics as predicted by LEB and LIB and example word-tags.

	% Women	% Black or African American
Bartender	56.1	7.4
Firefighter	3.5	6.8
Police officer	14.1	12.0

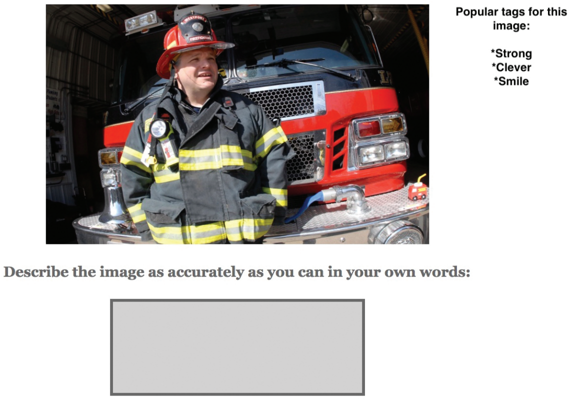
Table 1: % Total employed persons by gender and race

Table 1 details the participation of women and black or African Americans in the three professions, according to the 2016 U.S. Labor Force Statistics from the Current Population Survey[2]. These statistics reinforce the notion that images of white professionals will be more expected (i.e., stereotype confirming) across all professions, while images of black professionals of either gender will be less expected. In addition, women fighter fighters and police officers will be less expected than are men.

Task interface and manipulation. We implemented a very simple interface for the image-labeling task, modeled on ZenTag[3], a single-player tagging GWAP. Our interface displayed no information as to the purpose of the task or its designers, and had no incentive mechanisms beyond the payment. Figure 2 shows the condition in which social cues are provided to the participant, in the form of three "popular tags" for the given image. The tags included one concrete, positive attribute (smile) along with two abstract, positive characteristics (strong, clever), which were the same across images. As mentioned, this manipulation is designed to cue participants as to the communicative purpose of the task. The interface for the control setting was identical to that in Figure 2, but without the "popular tags."

[2] https://www.bls.gov/cps/cpsaat11.htm

[3] https://tiltfactor2.dartmouth.edu/zentag



Popular tags for this image:

*Strong
*Clever
*Smile

Describe the image as accurately as you can in your own words:

Figure 2: Task interface with social cues.

Procedure. We used a between-subjects design and recruited workers through MTurk. Participants were U.S. residents, had not participated in any previous tasks posted by our research group (and were subsequently entered in our list of ineligible workers), and were paid $1.00 per image annotated. We set up four HITs (2 social cues settings x 2 desired genders of participants) per image. Each HIT directed participants to the respective annotation task. We used MTurk's "premium qualification" function to recruit a gender-balanced set of participants, but also confirmed its accuracy by collecting participant responses. Before beginning, participants were asked to state their gender, ethnic group, and native language. In the current study, we analyze the data of participants who self-reported as being Caucasian and native English speakers (a total of 624 images labeled by men participants, and 636 images described by women). The full dataset is accessible at GitHub[4].

Analyzing textual descriptions. We used both automated and manual approaches in analyzing the language used to describe images. First, we used the Linguistic Inquiry and Wordcount (LIWC) (Pennebaker et al. 2015). For an input text, LIWC returns the percent of words that map onto a number of dictionaries consisting of conceptually meaningful words. With LIWCs default dictionary, we gauged the overall length of participants' descriptions (i.e., wordcount – WC) as well as their use of long words (i.e., 6+ letters – Sixltr). We use these measures to benchmark workers' general behaviors when describing images.

We also created a custom LIWC dictionary, based on the Subjectivity Lexicon (Wilson, Wiebe, and Hoffmann 2005). This is a resource of over 8.000 words, rated as to their type (i.e., whether they are strongly, or always, subjective, regardless of context, or weakly subjective) and polarity (i.e., conveying positive or negative sentiment). Our dictionary consists only of the strongly subjective words; examples are provided in Table 2. LIWC provides a continuous feature score (% words in the input description that are subjective).

	Adjective	**Noun**	**Verb**
Positive	pretty	smile	admire
Negative	ugly	coward	fear

Table 2: Example subjective words with polarity.

Question	**Example description**	κ
Does the description describe the person's physical **appearance**, beyond his/her gender and profession?	Black female firefighter in her gear holding a hose in her gloved hands (1,0,0)	0.88
Does it describe his/her **character** or **mood**?	A young, pretty, dark-haired female bartender who is friendly, lively, and lots of fun to be around (1,1,1)	0.89
Does the description cast **judgment** on the person in any way (bad or good)?	Arrogant, takes job seriously, very family oriented (0,1,1)	0.85

Table 3: Manual coding of descriptions with Fleiss' κ (Fleiss and Cohen 1973).

We also evaluated descriptions of images manually, for markers of abstraction and concreteness. Three independent analysts, unaffiliated with our research group, were asked to read each description, answering each of the yes/no questions shown in Table 3. Examples are shown along with their respective assessments for two markers of abstraction (character, judgment) and one marker of concreteness (appearance).

Table 4 examines correlations between the LIWC scores (subjectivity) and manual analyses of descriptions (markers of abstractness/concreteness)[5]. Descriptions that reference physical appearance, a concreteness marker, contain fewer subjective words, as compared to those that do not mention appearance. In contrast, descriptions that reference a person's character traits or mood, or that cast judgment, tend to contain more subjective words of both positive and negative polarities. In other words, descriptions deemed less subjective by the LIWC measure tend to convey concrete details (e.g., someone's clothing, how she is positioned), while descriptions deemed more subjective tend to refer to characteristics that must be inferred or projected by the annotator, such as the mood or character traits of the depicted person, or whether she is good at her job. These observations are consistent with previous work (Semin and Fiedler 1991), suggesting that our methods for analyzing descriptions are appropriate.

Hypotheses

Having reviewed the relevant literature and presented our methodology, we detail the three hypotheses to be tested:

H1: Linguistic Expectancy Bias (LEB). The LEB predicts that (1) white professionals will be described more abstractly than blacks (H1a), and (2) men will be described

[4] https://github.com/jahna/HCOMP2018

[5] We use the following indications of statistical significance: ***p<.001, **p<.01, *p<.05.

	Appearance	t	Character/Mood	t	Judgment	t
Subjective	4.6 / 2.5		12.4 / 7.9		12.3 / 9.1	
	9.6 / 3.7	-4.6^{***}	1.9 / 0	31.9^{***}	5.4 / 0	16.4^{***}
Positive	2.9 / 0		8.4 / 4.2		9.2 / 5.9	
	6.7 / 0	-4.7^{***}	1.1 / 0	21.9^{***}	3.2 / 0	14.4^{***}
Negative	1.6 / 0		3.8 / 0		2.8 / 0	
	2.8 / 0	-1.2	0.64 / 0	12.3^{***}	2.0 / 0	3.3^{**}

Table 4: Mean/median subjective, positive and negative words by manual assessments on characteristics (yes-top, no-bottom).

more abstractly than women (H1b), with the exception of bartenders, where the gender distribution is nearly 50/50.

H2: Linguistic Intergroup Bias (LIB). The LIB predicts that (1) white men will describe other white men more abstractly as compared to other social groups (H2a), and (2) white women will describe other white women more abstractly than other groups (H2b).

H3: Effect of social cues ("popular tags"). Since the presence of social information gives communicators a hint as to the purpose of the task, it is likely to trigger biases. We expect to observe that biases are more frequent in the social cues setting (H3).

Analysis

We first confirmed that participants approached the task in a similar manner across all images. We used ANOVA to test for differences in description length (log of the number of words written), by participant gender and by image, as well as their interaction. Only the main effect on gender was significant (F=15.324, $p < .001$). Men created longer descriptions of images (mean/median 18.3/16.5 words) as compared to women (15.6/13.0 words). However, there were no significant differences by participant gender or image with respect to the use of long words (i.e., the ANOVA with Sixltr as the response variable had no significant effects).

Testing for LEB

We fit statistical models to examine potential sources of variance in each of the three dependent variables, which are indications of abstractness in descriptions: 1) the use of subjective words (both positive and negative), 2) mentioning the target person's character/mood, and 3) making judgment in some way. We examined whether worker gender (G in the tables), the gender (ImG) and race (ImR) of the person depicted in the image, and all interactions, are related to abstraction in the descriptions of an image. For the continuous variable (percentage of words that are subjective), we used ANOVA with three main effects and four interaction terms. In case of significance, we provide effect sizes (η^2) (Cohen et al. 2013) and follow up with Tukey's Honestly Significant Differences (HSD) test (Williams and Abdi 2010) to better understand the sources of variance. Tukey's HSD is specifically used to compare group means in an ANOVA and does not require correction for multiple comparisons (Day and Quinn 1989).

For the discrete dependent variables, we fit logit regression models, to predict the likelihood of an image description containing each feature (referencing the perceived character/mood or casting judgment on the person), given the explanatory variables and all interactions. In the case of a significant coefficient, we provide the odds ratio as a measure of effect size (Hilbe 2011).

Testing for LIB

To examine for evidence of LIB, we separated the observations into cases where participants described images of in-group members versus other individuals (i.e., men participants describing the three images of white men professionals versus the other nine images; women participants describing the three images of white women professionals versus the other images). We then compared the two groups of observations with respect to the three independent variables mentioned above. For the continuous variable, we examine for differences across the groups of images using the Welch two-sample t-test. For the comparisons on the discrete variables, we use the Chi-squared test for equality of proportions.

Results: Control Setting

Table 5 details the model for the use of abstract language in descriptions provided by the workers in the task with no social cues, while Tables 7 and 9 detail the models for predicting references to character or mood, or for casting judgment on the person depicted in the image, respectively.

LEB. We find supporting evidence for H1a, which concerns race-based bias. However, this evidence is limited to the images of bartenders. Specifically, in the ANOVA, there is a significant main effect on the race of the person depicted (ImR), such that white bartenders are described with more subjective words than are the black bartenders, regardless of the gender of the participant or the gender of the depicted bartender. In addition, in the logit model to predict judgmental descriptions (Table 9), we observe that workers were more likely to be judgmental of the white woman (i.e., making interpretations about her), and less likely to judge the black woman bartender. There is no support for H1b, which predicts gender-based bias in the descriptions of the images. Neither the main effect on the gender of the depicted professional nor the interaction terms are significant in any model.

LIB. To evaluate H2a, we consider the image descriptions contributed by men, comparing the descriptions of in-group members versus others. There are no significant differences between groups in the mean use of subjective words ($t = -0.67$, $p > .05$). Likewise, there are no differences in references to character or mood ($\chi^2 = 0.26$, $p > .05$) or

	Gender	ImG	ImR	G*ImG	G*ImR	ImG*ImR	G*ImG*ImR	Sig. Diff
Bartender	.12	.79	12.9*** (.06)	.22	.10	.26	.13	ImR: White>Black
Firefighter	1.4	.07	.44	.90	.07	1.01	.03	n.s.
Police officer	.01	2.4	1.7	.89	1.0	.43	.98	n.s.

Table 5: Control setting - Use of strongly subjective words in image descriptions (ANOVA F and η^2).

	Gender	ImG	ImR	G*ImG	G*ImR	ImG*ImR	G*ImG*ImR	Sig. Diff
Bartender	22.3*** (.08)	8.3** (.06)	15.9*** (.06)	.15	.00	4.2* (.02)	.92	G: Women>Men ImG: Men>Women ImR: White>Black ImG*ImR: WM,WW,BM>BW
Firefighter	5.8* (.02)	5.2* (.02)	.94	.55	.30	2.65	.50	G: Women>Men ImG: Men>Women
Police officer	9.9** (.04)	.01	.55	.21	2.7	.29	1.9	G: Women>Men

Table 6: Social cues setting - Use of strongly subjective words in image descriptions (ANOVA F and η^2).

	Gender	ImG	ImR	G*ImG	G*ImR	ImG*ImR	G*ImG*ImR
Bartender	.85	-.31	-1.1	-.49	-1.1	-1.2	1.1
Firefighter	-.55	-.47	-.39	-.04	.45	.22	.03
Police officer	-.45	.46	1.1	.26	-.31	-.15	-.06

Table 7: Control setting - Descriptions that reference the character/mood of depicted person (logit t and odds ratio).

in casting judgment ($\chi^2 = 3.59$, $p > .05$). To evaluate H2b, we compare the descriptions contributed by white women on the images of white women versus others. No significant difference is detected for the extent to which subjective words are used ($t = -.07$, $p > .05$). Likewise, no significant differences are detected with respect to the manually evaluated characteristics: character/mood ($\chi^2 = .196$, $p > .05$), casting judgment ($\chi^2 = .007$, $p > .05$).

Results: Social Cues Setting

LEB. We first evaluate the models concerning the use of subjective words in image descriptions, using the interface in which social cues were provided (Table 6). Here, we observe that, in contrast to the control setting, there is a significant main effect for participant gender (G), with women using more subjective words when describing all three sets of images, as compared to men. In addition, for images of bartenders and firefighters, there is evidence of gender-based bias (H2b), with images of men being described more abstractly (i.e., with more subjective words) as compared to images of women. For bartenders, race-based bias is again observed, as the main effect on the race of the depicted person (ImR) is significant (H1a).

The analysis on whether participants inferred aspects of the depicted person's character or mood (Table 8) also provides evidence for gender- and race-based LEB. The negative coefficients for ImG and ImR in models for bartender and firefighter indicate that participants were less likely to describe the character or mood of women as compared to men, and of African Americans as compared to white professionals.

LIB. We also test H2a, in the social cues setting. The t-test on the continuous variable reveals that white men use more subjective words when describing images of other white men, compared to other social groups ($t = 3.69$, $p < .001$). There is no significant difference with respect to making judgments of depicted individuals ($\chi^2 = 1.33$, $p > .05$). However, white men did describe the character or mood of images of white men more often than they did images of the other social groups ($\chi^2 = 17.6$, $p < .001$).

In contrast, we find no evidence to support H2b. White women were not more likely to use positive, subjective words when describing images of white women in comparison to others ($t = 1.10$, $p > .05$), nor were they more likely to infer aspects of white women's characters or moods ($\chi^2 = .22$, $p > .05$) or to judge them ($\chi^2 = .28$, $p > .05$) more often than others.

Discussion

To foster better understanding of the possible sources of bias in human-generated data, we evaluated a HIT based on visual perception that is often crowdsourced - annotating images depicting people of different genders and races, through textual descriptions (i.e., free text). We intentionally provided vague instructions and little contextual information in the control setting, in order to see how contributors would proceed on task. In contrast, our test setting included "popular tags for this image" to serve as social cues, which subtly provide contributors with information concerning the task's intended purpose, as well as audience (i.e., for whom the descriptions are intended).

Table 11 summarizes the linguistic biases, by type, social construct, and worker race/gender (in the case of LIB), which were observed in the descriptions produced by workers in the two settings. Even in the control setting, workers produced descriptions for images of bartenders that ex-

	Gender	ImG	ImR	G*ImG	G*ImR	ImG*ImR	G*ImG*ImR
Bartender	.07 .	-3.0** (.70)	-2.5* (.75)	2.2* (1.43)	.85	.40	-1.0
Firefighter	.31	-2.9** (.68).	-1.9* (.77)	1.2	-.02	1.3	-.04
Police officer	-.11	-.48	-.85	.81	2.6* (1.16)	.85	-1.96

Table 8: Social cues setting - Descriptions that reference the character/mood of depicted person (logit t and odds ratio).

	Gender	ImG	ImR	G*ImG	G*ImR	ImG*ImR	G*ImG*ImR
Bartender	-.37	2.2* (1.24)	.97 .	-.94 .	.32 .	-2.5* (0.71)	.82
Firefighter	.04	.09	.34	.01	-.27	.30	.20
Police officer	.59	.46	1.8	-.35	-1.8	-.33	.87

Table 9: Control setting - Descriptions that cast judgment on depicted person (logit t and odds ratio).

	Gender	ImG	ImR	G*ImG	G*ImR	ImG*ImR	G*ImG*ImR
Bartender	.33	-.75	-.10	.75	.20	.54	-1.08
Firefighter	1.2	.42	-1.3	-.89	-.31	.56	.53
Police officer	.68	-.56	-.86	.49	2.4* (1.50)	.04	-.54

Table 10: Social cues setting - Descriptions that cast judgment on depicted person (logit t and odds ratio).

Type of bias	**Control**	**Social Cues**
LEB	Race (H1a)	Race (H1a) Gender (H1b)
LIB - WM	None	Positive subjective words Character / mood (H2a)
LIB - WW	None	None

Table 11: Summary of hypotheses supported.

hibited race-based linguistic biases, which can be attributed to the LEB. We are not immediately able to explain why this did not also occur for images of police officers and firefighters. It can be noted that the bartender images depicted individuals working in an upscale bar/restaurant; thus, it is possible that this occupation was not interpreted as being as working class as the others (police officer and firefighter).

In the test condition, in which workers were exposed to three positive, subjective "popular tags" for an image, linguistic biases based on both gender and race were observed. As predicted by the LEB, images of women and individuals of color were described less abstractly than images of white male professionals.

Furthermore, as predicted by theory, evidence of in-group biases (i.e., LIB) was observed only in the test condition. Here, white men contributors provided more abstract descriptions of the images of the three white men professionals, as compared to all other images. Interestingly, the descriptions produced by white women workers did not exhibit evidence of LIB. This is a particularly interesting finding given that, as seen in Table 6, women are generally more likely than men to use subjective words when describing images (i.e., there is a main effect on participant gender, for all image sets in the test setting). Despite this, women's descriptions on the images do not exhibit LIB; they are not more likely to describe images depicting members of their own in-group, based on race and gender, more abstractly than others.

At a first glance, it might seem obvious that displaying "popular tags" of images, which consist of positive, subjective descriptive words, would necessarily trigger participants' use of more subjective words and inferential descriptions. After all, there is a good deal of empirical evidence that human communicators mimic one another's linguistic styles (e.g., (Giles and Powesland 1997; Ireland et al. 2011)). However, what is striking is that the increase in the use of abstract, subjective descriptions is not uniform across all images, but rather is consistent with theories of linguistic bias.

Limitations

Our study focused on one specific HIT, in which we incorporated social cues in a particular way, by showing "popular tags" on an image. Our popular tags were chosen by the researchers to have specific properties - they were positive, subjective words - and were not actually assigned by other workers. There are, of course, many different ways in which we could provide workers with context on a task, depending on the goals and type of annotation desired. Given the interdependencies between HIT design decisions, we should be careful not to assume that others will necessarily obtain the same results that we did. That said, the extent to which our results were predicted by theory was striking.

Furthermore, we have limited information concerning the crowdworkers who participated in our task. In addition, due to our between-subjects design, we did not test whether the same contributor's descriptions across all images were consistently in line with theory, which would provide even stronger evidence of the influence of linguistic biases. Nonetheless, the between-subjects design helps to prevent learning effects and/or boredom on task, which might have occurred if we had asked each contributor to describe all 12 images.

Conclusions

Linguistic biases present a challenge for the generation of quality annotations for multimedia depicting people. A key take-away from the present study is that the task we examined - free-text annotation of images - is fundamentally a communication process. Images embody rich but implicit information about the subject(s) they depict; thus, the worker is asked to make this implicit information explicit, through the choice of words that describe the image. However, while language facilitates the expression of information we intend to convey, it also reflects the social stereotypes that influence the way we perceive others.

Sociolinguists have long argued that demographic characteristics, and in particular gender, are strong influences on people's language patterns (Labov 1990). Indeed, we found systematic differences between responses of women and men contributors, which is in line not only with theory but also other recent findings on the crowdsourcing of image annotation. For instance, Matsangidou and colleagues (Matsangidou et al. 2018) asked workers to describe images of people depicted in painful situations, as well as to evaluate their own levels of emotional arousal. While men and women reported similar levels of arousal when viewing the images, as well as comparable levels of confidence in the accuracy of their work, significant qualitative differences in the words used to describe the images were reported.

As linguistic bias is population-wide, HIT owners will need to decide how to address it. They will also need to carefully tune the design of an annotation HIT for people-related images. The present work has particularly highlighted the role of social cues, in potentially swaying workers' responses. Future research should continue to examine the trade-offs between the provision of social information, which contains cues about the purpose of the data to be provided by workers, and the manifestation of linguistic biases.

Acknowledgments

This project is partially funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 739578.

References

Ames, M., and Naaman, M. 2007. Why we tag: motivations for annotation in mobile and online media. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 971–980. ACM.

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.

Cohen, J.; Cohen, P.; West, S. G.; and Aiken, L. S. 2013. *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge.

Day, R., and Quinn, G. 1989. Comparisons of treatments after an analysis of variance in ecology. *Ecological monographs* 59(4):433–463.

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*.

Fleiss, J. F., and Cohen, J. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement* 33(3):613–619.

Giles, H., and Powesland, P. 1997. Accommodation theory. In *Sociolinguistics*. Springer. 232–239.

Hilbe, J. M. 2011. Logistic regression. In *International Encyclopedia of Statistical Science*. Springer. 755–758.

Ipeirotis, P. G.; Provost, F.; and Wang, J. 2010. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation*, 64–67. ACM.

Ireland, M. E.; Slatcher, R. B.; Eastwick, P. W.; Scissors, L. E.; Finkel, E. J.; and Pennebaker, J. W. 2011. Language style matching predicts relationship initiation and stability. *Psychological science* 22(1):39–44.

Kamar, E.; Hacker, S.; and Horvitz, E. 2012. Combining human and machine intelligence in large-scale crowdsourcing. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, 467–474. International Foundation for Autonomous Agents and Multiagent Systems.

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*.

Kazai, G.; Kamps, J.; and Milic-Frayling, N. 2012. The face of quality in crowdsourcing relevance labels: Demographics, personality and labeling accuracy. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, 2583–2586. New York, NY, USA: ACM.

Kazai, G.; Kamps, J.; and Milic-Frayling, N. 2013. An analysis of human factors and label accuracy in crowdsourcing relevance judgments. *Information Retrieval* 16(2):138–178.

Labov, W. 1990. The intersection of sex and social class in the course of linguistic change. *Language variation and change* 2(2):205–254.

Law, E., and von Ahn, L. 2009. Input-agreement: a new mechanism for collecting data using human computation games. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1197–1206. ACM.

Law, E., and von Ahn, L. 2011. Human computation. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 5(3):1–121.

Maass, A.; Salvi, D.; Arcuri, L.; and Semin, G. R. 1989. Language use in intergroup contexts: The linguistic intergroup bias. *Journal of personality and social psychology* 57(6):981.

Manzo, C.; Kaufman, G.; Punjasthitkul, S.; and Flanagan, M. 2015. " by the people, for the people": Assessing the value of crowdsourced, user-generated metadata. *DHQ: Digital Humanities Quarterly* 9(1).

Matsangidou, M.; Otterbacher, J.; Ang, C. S.; and Zaphiris, P. 2018. Can the crowd tell how I feel? Trait empathy and ethnic background in a visual pain judgment task. *Universal Access in the Information Society* 1–13.

Park, J.-R. 2009. Metadata quality in digital repositories: A survey of the current state of the art. *Cataloging & classification quarterly* 47(3-4):213–228.

Pennebaker, J. W.; Boyd, R. L.; Jordan, K.; and Blackburn, K. 2015. The development and psychometric properties of liwc2015. Technical report.

Rader, E., and Wash, R. 2008. Influences on tag choices in del. icio. us. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, 239–248. ACM.

Ross, J.; Irani, L.; Silberman, M. S.; Zaldivar, A.; and Tomlinson, B. 2010. Who are the crowdworkers?: Shifting demographics in mechanical turk. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '10, 2863–2872. New York, NY, USA: ACM.

Semin, G. R., and Fiedler, K. 1991. The linguistic category model, its bases, applications and range. *European review of social psychology* 2(1):1–30.

Semin, G. R.; de Montes, L. G.; and Valencia, J. F. 2003. Communication constraints on the linguistic intergroup bias. *Journal of Experimental Social Psychology* 39(2):142–148.

von Ahn, L., and Dabbish, L. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 319–326. ACM.

von Ahn, L., and Dabbish, L. 2008. Designing games with a purpose. *Communications of the ACM* 51(8):58–67.

Wigboldus, D. H.; Semin, G. R.; and Spears, R. 2000. How do we communicate stereotypes? linguistic bases and inferential consequences. *Journal of personality and social psychology* 78(1):5.

Wigboldus, D. H.; Semin, G. R.; and Spears, R. 2006. Communicating expectancies about others. *European Journal of Social Psychology* 36(6):815–824.

Williams, L. J., and Abdi, H. 2010. Post-hoc comparisons. *Encyclopedia of Research Design* 1060–1067.

Wilson, T.; Wiebe, J.; and Hoffmann, P. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, 347–354. Association for Computational Linguistics.

Winkielman, P.; Halberstadt, J.; Fazendeiro, T.; and Catty, S. 2006. Prototypes are attractive because they are easy on the mind. *Psychological science* 17(9):799–806.

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 Conference on Empirical Methods in Natural Language Processing (EMNLP).*

Zhuang, H., and Young, J. 2015. Leveraging in-batch annotation bias for crowdsourced active learning. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, 243–252. New York, NY, USA: ACM.