What Makes an Image Tagger Fair?

Proprietary Auto-tagging and Interpretations on People Images

Pınar Barlas
Kyriakos Kyriakou
Research centre for Interactive media,
Smart systems and Emerging technologies
Nicosia, CYPRUS

Styliani Kleanthous
Jahna Otterbacher
Cyprus Center for Algorithmic Transparency
Open University of Cyprus
Nicosia, CYPRUS

ABSTRACT

Image analysis algorithms have been a boon to personalization in digital systems and are now widely available via easy-to-use APIs. However, it is important to ensure that they behave fairly in applications that involve processing images of people, such as dating apps. We conduct an experiment to shed light on the factors influencing the perception of "fairness." Participants are shown a photo along with two descriptions (human- and algorithm-generated). They are then asked to indicate which is "more fair" in the context of a dating site, and explain their reasoning. We vary a number of factors, including the gender, race and attractiveness of the person in the photo. While participants generally found human-generated tags to be more fair, API tags were judged as being more fair in one setting - where the image depicted an "attractive," white individual. In their explanations, participants often mention accuracy, as well as the objectivity/subjectivity of the tags in the description. We relate our work to the ongoing conversation about fairness in opaque tools like image tagging APIs, and their potential to result in harm.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision tasks**; • **General and reference** → *Experimentation*; • **Information systems** → *Personalization.*

KEYWORDS

algorithmic bias, computer vision, fairness, image analysis

ACM Reference Format:
Pınar Barlas, Kyriakos Kyriakou, Styliani Kleanthous, and Jahna Otterbacher. 2019. What Makes an Image Tagger Fair?: Proprietary Auto-tagging and Interpretations on People Images. In *27th Conference on User Modeling, Adaptation and Personalization (UMAP '19), June 9–12, 2019, Larnaca, Cyprus.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3320435.3320442

1 INTRODUCTION

Image recognition is clearly one of the success stories of modern machine learning. Since Krizhevsky and colleagues [13] first applied deep learning to their entry in the ImageNet Challenge,[1] the technology has made rapid progress. Beyond early applications in restricted domains (e.g., processing satellite imagery), image analysis algorithms are now widely used in commercial applications and social media, enabling functionality that users take for granted, such as the ability to search and retrieve images in real time, based on content - even in the absence of descriptive metadata.

The increased performance of image analysis algorithms is proving to be a boon to technologies where user modeling, personalization and adaptation are required. For instance, they are said to be transforming retail.[2] In e-stores, image recognition is used to curate and recommend "personal styles" for a given shopper, by recognizing the visual characteristics of items he or she has viewed to date and building a user model around those.[3] During in-store shopping, image recognition is used to understand where the user is positioned or what interests her, to make recommendations or even to trigger an enhanced "augmented reality product experience."[4]

One of the leading providers of image analysis algorithms, Clarifai, reports that the technology is also used extensively in the context of dating apps.[5] For example, an app called "Hinge" is tracking user behavior, using algorithms to determine who and what users liked while interacting with the app, to act as a "visual matchmaker." Another client used Clarifai's image analysis algorithms to help dating app users better craft their profile images.[6]

In short, image analysis technology is having a growing influence on our digital interactions and experiences. However, at the same time, there is awareness that the technology has some serious ethical concerns. For example, in December 2017, Apple announced refunds to Chinese users of the iPhone X, after complaints that its Face ID technology could not distinguish between Asian faces.[7] In a similar vein, three years after a blunder by Google Photos, in which a Black user's image was labeled with a racist tag, the company announced its promised solution. However, the "fix," which was to simply remove the offending tag from the database, was highly criticized as being a "workaround," rather than a true solution.[8]

There is a growing literature surrounding the behaviors of computer vision algorithms, and their potential to treat people unfairly. Zhao and colleagues [26] documented evidence of gender-based

[1] http://www.image-net.org/challenges/LSVRC/

[2] https://www.forbes.com/sites/forbesagencycouncil/2018/07/16/use-ai-to-create-a-more-personalized-profitable-customer-experience/#77560595f3a7

[3] https://vue.ai/solutions/omnichannel-personalization.html

[4] https://catchoom.com/blog/image-recognition-enables-scan-to-shop-retail-experiences/

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

[6] https://blog.clarifai.com/clarifai-featured-hack-use-ai-to-tune-up-your-online-dating-profile

[7] https://www.newsweek.com/iphone-x-racist-apple-refunds-device-cant-tell-chinese-people-apart-woman-751263

[8] https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai

UMAP '19, June 9–12, 2019, Larnaca, Cyprus

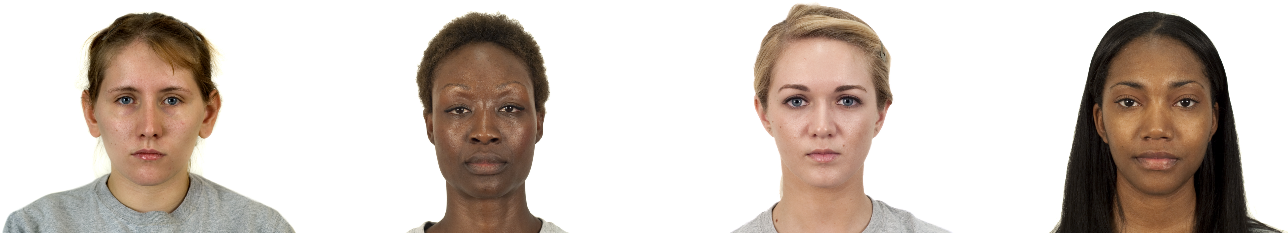

Figure 1: Chicago Face Database (CFD) [14] images of average-looking (WF-036,BF-231) and attractive women (WF-233,BF-233).

Tag source	Clarifai	Crowdworkers from US
BF-231	(adolescent,) (adult,) Afro, athlete, casual, (child,) face, facial, isolated, (man,) one, pensive, people, person, portrait, profile, side, wear, (woman)	African American, black, curly, eyes, full, frizz, lips, (middle aged,) serious, shiny skin, short hair, strong, (woman)
WF-036	(adolescent,) casual, (child,) contemporary, cute, eye, facial, fashion, fine looking, fun, funny, isolated, looking, (man,) one, people, portrait, serious, wear, (young)	blue eyes, brown hair, caucasian, front view, (girl,) lip gloss, long hair, plain expression, round face, short bangs, sober, solo, white background, (woman,) (young)
BF-233	casual, cute, eye, facial expression, fashion, isolated, look, looking, one, pensive, people, portrait, pretty, serious, wear	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
WF-233	casual, cute, eye, fashion, fine-looking, friendly, hair, isolated, joy, look, looking, one, people, portrait, pretty, serious	attractive, bigger ears, blonde, blond streaks, blue eyes, grey shirt, lip mole, mouth, nice eyebrows, pale skin, sandy hair, serious, small lips, t-shirt, thin lips, white
WM-004	casual, cute, eye, face, facial expression, fashion, fine-looking, friendly, hair, isolated, look, looking, one, pensive, people, portrait, wear	angry, brown hair, clean shaven, eyes, green eyes, grey sweatshirt, lips, long hair, messy hair, neck, nose, shirt, small ears, small lips, thick hair
BM-234	casual, desktop, face, facial expression, fine-looking, friendly, isolated, look, looking, one, pensive, people, portrait, satisfaction, serious, studio, wear	average skin, average lips, black hair, brown eyes, bushy eyebrows, happy expression, latin, serious, short hair, smile, stubble, tired, t-shirt, trusted
BM-009	casual, cool, face, facial expression, fine-looking, friendly, happiness, indoors, isolated, look, looking, one, pensive, people, portrait, satisfaction, wear	adam's apple, African American, chin, dark, dark eyebrows, dark eyes, big lips, black hair, brown eyes, emotionless, grey shirt, mugshot, not smiling, short hair
WM-022	casual, cool, eye, face, fashion, fine-looking, friendly, hair, isolated, look, looking, one, people, portrait, serious, studio, wear	ears, eyes, face, grey shirt, hair, head, mouth, nose, round face, shadow, sleepy looking, stubble beard, thick eyebrows, white skin

Table 1: Output tags for CFD images produced by Clarifai image analysis service as well as crowdworkers. Tags referring to gender and age are in parentheses, and were removed from the list after the first pilot study.

bias in the popular MS-COCO dataset. They found that labels describing activities depicted in images were highly gendered in a stereotypical way (e.g., verbs such as "cooking" or "shopping" were associated with images of women). Another study cited 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 between error rates can be more than 30% [6]. Similarly, Rhue [20] reported biases in emotion tagging, with Black men being more likely to be tagged with a negative emotion than White men, when using Face++ and Microsoft's Face API.[9]

It is crucial to understand how image analysis algorithms treat people-related media, and in turn, how users interpret their behaviors. This is particularly important because the use of proprietary algorithms by third party developers is on the rise, through their commercialization, a phenomenon Gartner has called the "Algorithm Economy."[10] In short, image tagging algorithms, such as the above-mentioned, are available to developers as software-as-a-service. They are easy to use, although opaque; the algorithm we use in our study, the Clarifai API,[11] does not provide developers with the set of possible descriptive tags that it uses to describe an input image. Thus, its "social behaviors" are unpredictable.

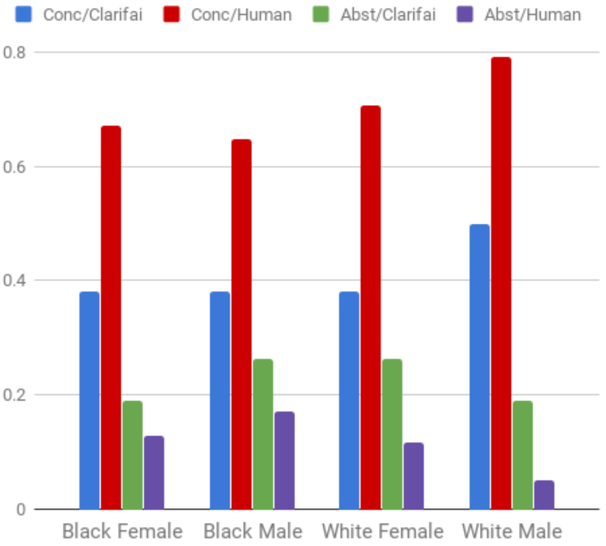

Figure 2: Mean proportion of concrete/abstract tags on images in our experiments, by race/gender, and tag source.

We are not aware of previous work that examines *users' perceptions* of fairness when it comes to image analysis algorithms'

[9] https://azure.microsoft.com/en-us/services/cognitive-services/face/
[10] https://www.gartner.com/smarterwithgartner/the-algorithm-economy-will-start-a-huge-wave-of-innovation/
[11] https://clarifai.com

interpretations of people images. As illustrated in Table 1, which shows human- and algorithm-generated descriptions on four photos (Figure 1), algorithms often go beyond describing concretely observed attributes, making inferences as to a depicted individual's character traits or judging her physical attractiveness. This is illustrated in Figure 2, which shows the distribution of concrete/abstract tags, used by Clarifai and human "taggers," in describing the images used in our study. When taggers use words that are too inferential, there is a danger that users might perceive them as behaving in ways that are not socially just. We conduct an experiment at Amazon Mechanical Turk[12] to address three research questions:

1. Which description (human- or algorithm-generated) is perceived as being "more fair"?

2. What are the key factors users consider when judging a tagging algorithms' treatment of people images?

3. How do characteristics of the depicted person - such as gender, race, and physical attractiveness - and the participant's gender impact the perception of algorithmic fairness?

2 BACKGROUND

To ground our study, we first establish the centrality of users' visual self-presentation in their online interactions. We then discuss current research in the emerging area of fairness, accountability and transparency (FAT) in algorithmic systems and processes. Finally, we examine the potential role of image analysis algorithms in personalization and user adaptation in dating applications.

2.1 Visual self-presentation

The role of physical appearance in human interaction cannot be denied; after all, it is the first characteristic that others observe in a social interaction [2]. Thus, it is not surprising that users' visual self-presentation is important to them. Media psychologists have explained that the practice of uploading "selfies" is related to one's feelings of self-worth [23]. More generally, there are findings to suggest that social media profiles are projections of the idealized self [7]. At the same time, some have argued that the media culture's focus on appearance and users' repeated exposure to this is correlated to increased body image disturbance [15]. There is also evidence that the "what is beautiful, is good" stereotype applies in online interactions as well as those face-to-face. Brand and colleagues [5] found that individuals with attractive profile photos in dating websites, are viewed more favorably and with more positive qualities, as compared to those with less attractive photos.

The prevalence of algorithmic processes in social spaces has complicated the process of self-presentation [8], and recent research cites that users want more information - and control - in managing how algorithms profile them and/or mediate in their presentation [1]. Given the above, one can envision how automated image analysis could adversely affect users' well-being when output tags on images are offensive or otherwise seen as unjust. Therefore, it is crucial to understand how algorithms interpret people-related media and in turn, how users themselves judge algorithmic behaviors.

2.2 Fairness, accountability and transparency

There is intense interest in the social side of algorithmic behaviors, and how to detect and redress their biases. Attention to the issue no doubt stems from the influence of opaque, proprietary algorithms in our information ecosystem. Diakopoulos describes algorithms as "power brokers" that are not always held accountable for their actions [9]. They are increasingly delegated everyday tasks and operate autonomously, with minimal human intervention [24]. Human behavior tends to reinforce algorithmic power and autonomy; there is a tendency for users to perceive them as objective [11, 16] while some remain totally unaware of algorithmic interventions in the systems they use [10]. Automated content analysis on images is an example of an "everyday" task which, as mentioned, underlies many applications and personalization mechanisms.

Researchers are developing auditing processes to "open the black box" in an effort to make algorithms more transparent to users and promote fairness [22]. However, in many contexts, it is difficult to define and operationalize a notion of "algorithmic fairness." As noted by Binns [4], fairness "is best understood as a placeholder term for a variety of normative egalitarian considerations."

Image tagging algorithms do not have an obvious baseline for comparison, in contrast to other processes. For instance, Kay and colleagues [12] compared Google image search results on professions (e.g., a search for "doctor" versus "nurse") with respect to the gender distribution of people depicted in the retrieved images. In this context, they used offline labor statistics to measure the deviation from what might be expected if the search engine provided an unbiased reflection of society. In the case of image analysis algorithms' descriptions on people images, it is difficult to say which words (i.e., tags) we should expect, or how we might determine if the observed behavior is fair. To this end, we aim to shed light on the dimensions of tagger behavior that users consider, when we ask them to explain whether descriptions of people images are fair. Furthermore, we shall correlate our findings to the properties of the images being analyzed, and in particular, the demographic and physical characteristics of the individual depicted.

2.3 Dating apps and personalized recommendations

The user modeling and recommender systems communities have long been looking into techniques and algorithms for recommending items or people to a user based on their user model; such techniques have also been applied in reciprocal systems. In reciprocal systems, where the receiver of the recommendations can be more than one user (e.g., dating systems, recruitment systems), the "fairness" aspect of the information that is implicitly assigned to user profiles must be carefully considered. The recommendations that are produced affect not only the receiver but also the person whose profile is in the recommendation. Of particular relevance, online dating services are using both explicit and implicit [19] user modeling for developing recommendations for possible matches by tracking user behavior in the system. Pizzato et al. [17] examined the interactions between users in a dating site (e.g., message exchange, profile viewing) and found that implicit preferences produced better recommendations compared to explicit preferences provided by users.

[12] http://www.mturk.com

However, the challenge in reciprocal recommender systems is that the recommendations do not target a single person but need to consider the preferences of other users with whom this person is interacting [18]. Zheng et. al. [27] performed an exploratory analysis on real data from a speed dating application, taking into account the user's expectations towards their recommendations with improved results compared to earlier approaches. Xia et al. [25] designed recommendations based on the interest similarity between two users if they send messages to the same users, and attractiveness similarity if they receive messages from same users in attempt to acknowledge the reciprocity of online dating recommender systems.

Recently, with the increasing accuracy of computer vision algorithms, apps have begun using image tagging APIs for recommendations and content organization through automated image tagging. For example, Architizer[13] is a marketplace that helps architects find the building products they need using Clarifai API to generate recommendations. Furthermore, as discussed in the introduction, Clarifai is used extensively in the context of dating apps. It can assist in implicitly matching people in an application or help the users better craft their profile image by tag recommendations. Hence, it is important to understand the perception of fairness in image tagging, especially when used in reciprocal systems. In this paper, we focus on the context of online dating services as a case study for identifying and analyzing people's perceptions of fairness, when they are presented with two sets of tags (one assigned by a human tagger and the other by an image tagging algorithm).

3 METHODOLOGY

We conducted a between-subjects experiment at MTurk. For each of our pilot studies and experiment, we recruited a gender balanced set of participants based in the United States. Specifically, MTurk's premium qualification function was used to ensure that each image in a given setting was analyzed by exactly 20 women and 20 men. U.S.-based participants were deemed to be the most appropriate, given that the image tagging technologies we are researching are provided by U.S. companies. In addition, participants were paid $1.00 for their time, with five minutes being the average time to completion. Finally, to avoid learning effects, participants could only complete one task in our study.

3.1 Images and descriptions

The images used in our study come from the Chicago Face Database (CFD). The CFD is a free resource[14] consisting of high-resolution, standardized images of diverse individuals, between the ages of 18 and 40 years, along with objective facial measurements and subjective norming data. Created by psychologists [14], the CFD is designed to facilitate research on a broad range of behavioral phenomena (e.g., social stereotyping and prejudice, interpersonal attraction). For our purposes, a significant benefit of using the CFD to study image tagging algorithms, is that the individuals are depicted in a similar, neutral manner, as shown in Figure 1. The CFD also contains subjective measurements on each image, collected from 30+ judges. For our study, we have relied on judges' perceptions of the physical attractiveness of the depicted individuals. Individuals referred to as "average-looking" had a mean score of 2.6 (out of 7) in the CFD, while "attractive" individuals had a mean score of 5.6.

The human- and algorithm-generated descriptions on images come from our publicly-available Social B(eye)as Dataset (SBD).[15] The dataset contains descriptions of all 597 CFD images, produced by six proprietary image tagging services including Clarifai, Google Vision API and others, as well as two sets of crowdworkers located in two large anglophone markets (the U.S. and India) [3]. In the current study, the algorithm-generated descriptions on people images that we present to participants are produced by Clarifai, while the human-generated descriptions were provided by U.S. based crowdworkers. In all cases, we provide the CFD image code (e.g., WF-036) so that our data sources can be traced through the CFD and SBD.

3.2 Pilot studies

To develop the experimental set-up, as well as the approach to analyzing the collected data, we conducted two pilot studies. Table 2 summarizes the sets of descriptive tags presented to participants, the target images used, as well as the genders of participants in the pilots. We used two photos of "average" women (WF-036, BF-231, pictured in Figure 1) as well as the descriptive tags provided by Clarifai and the human analysts (as shown in Table 1).

Table 2: Pilot studies.

	Tags	Image	Participant gender (W/M)
1	Unprocessed	BF-231	20/20
		WF-036	20/20
2	Corrected	BF-231	20/20
		WF-036	20/20

The instructions read as follows: "Today, many automated tools are used to generate descriptions of images on the Web. However, some tools exhibit biases when processing images of people. Given an image and two descriptions of its content, decide which one is **more fair**." After being presented with the image, plus the two descriptions, they were asked the following: "Imagine that you use auto-tagging in your personal photo collection. Which of the above descriptions is more fair? Enter 0 if you cannot tell." After entering their answer, they were prompted to "explain your answer regarding fairness." It should be noted that participants were not explicitly told that one description was machine- and one was human-generated.

3.2.1 *First pilot: unprocessed tags.*

As can be seen in Table 1, the Clarifai tagger often uses gender- and age-related tags inaccurately, in contrast to the human analysts. Perhaps unsurprisingly, the pilot revealed that participants largely felt that the human-generated descriptions were "more fair." Only five of 80 participants indicated that Clarifai's tags provided a more fair description of the depicted individual. This happened three times on BF-231 and twice for WF-036. Interestingly, all five of those participants were men.

[13] https://www.clarifai.com/customers/architizer

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

[15] Barlas, P., Kyriakou, K., Kleanthous, S. & Otterbacher, J. "Social B(eye)as Dataset", https://doi.org/10.7910/DVN/APZKSS, Harvard Dataverse, V1.

Participants' explanations of their answer were largely focused on the issue of *accuracy* with respect to the use of age- and gender-related tags. In all, 57 out of 80 responses discussed 'accuracy" , with 28 focusing on the demographic characterizations, particularly the incorrect use of the "man" and "child" tags.

3.2.2 Second pilot: corrected tags. In the second pilot, we revised the tags in our prompt, removing all gender- and age-related tags from both the human- and algorithm-generated descriptions, in order to enable participants to focus on deeper issues concerning the fairness of the taggers, and such that they did not simply equate fairness with accuracy. As expected, once the demographic tags were removed, there was an increase in the number of participants who found Clarifai's tags to be more fair (5 men and 5 women). In addition, the explanations of their answers went beyond discussing accuracy, with other themes emerging, such as the objectivity/subjectivity of the tags, as well as the extent to which tags were easily understandable.

3.2.3 Dimensions of fairness. In order to undercover the key factors that participants considered when judging the fairness of the descriptive tags, we first conducted a thematic analysis on the "fairness explanations" collected in the pilot studies, using an inductive approach. Two researchers analyzed the participants' free-text explanations independently to define the emerging dimensions discussed. Upon agreeing on the common dimensions, they independently sorted the responses from the second pilot into these categories. Their results were compared, the disagreements discussed, and sometimes a dimension's definition amended to come to a final consensus. In total, ten dimensions of fairness were discussed by participants, as described below. Responses (i.e., "fairness explanations") were subsequently coded for the presence/absence of these dimensions, which are not mutually exclusive.

Gender, **Race**, and **Age.** The participant mentions the depicted person's identity when explaining his or her answer. The responses that include the tags from the prompt referring to gender, race, or age were coded with the respective attribute(s). After the first pilot, tags referring to gender and age were removed from the prompt; however, participants occasionally brought up the absence of gender-related tags while Age was not mentioned.

"It does not emphasize racial characteristics."

Accuracy. The response discusses whether the tags are "correct," "factual," or similarly considered to correspond to the truth. This also includes responses which discuss if the tag is more general (e.g., "eye") or more specific (e.g., "blue eyes").

"Number 2 is fair as the description is more accurate."

Objectivity/Subjectivity. The explanation discusses whether the tags are based on "concrete" characteristics or if they make assumptions or embody opinions. Responses that provide examples of subjective tags but do not explicitly talk about their subjectivity are not considered here.

"Description 2 is more fair because it is not subjective and is accurate and less open to interpretation."

Physical Characteristics. The response discusses whether the tags are based on features of the person which can be directly observed. These responses point out specific tags discussing body parts or hair, or other features of the photo which are "visible."

"I liked that it focused on aspects about the image, such as her hair and eye color."

Biases. The explanation explicitly talks about the tags, or the system producing the tags, being socially biased.

"It's more descriptive and less biased."

Racist. The response explicitly talks about a racial bias in the tagging or tags which can be considered racist.

"I feel like it sounds the best without sounding racist in any way."

Political correctness. The responses coded with this dimension discuss how the tags may be perceived by the users. For example, the description may be "offensive", "rude", "mean", or "nice".

"A lot of the words would not be described as favorable or putting the person in a good light."

Understanding. The response discusses whether the tags would help a user understand what is in the image, give a "good description," or are "easy to understand."

"If someone gave me that I would be able to tell what the person looked like easier than description 1."

3.3 Experimental set-up

Informed by the findings of our pilot studies, we designed an experiment that asked participants to consider the descriptive tags in a more socially sensitive application. This time, they were told to "Imagine that auto-tagging is used to facilitate searching profiles of people at a dating site. In that context, which of the two descriptions is more fair? Enter 0 if you cannot tell." Table 3 summarizes the crowdwork experiment in terms of the factors that were varied.

Table 3: Experimental set-up.

Image	Race	Gender	Appearance	Participants (W/M)
BF-231	Black	Woman	Average	20/20
WF-036	White	Woman	Average	20/20
BF-233	Black	Woman	Attractive	20/20
WF-233	White	Woman	Attractive	20/20
BM-009	Black	Man	Average	20/20
WM-022	White	Man	Average	20/20
BM-234	Black	Man	Attractive	20/20
WM-004	White	Man	Attractive	20/20

4 ANALYSIS

We now analyze the 320 responses of participants, collected in the experiment. First, we explore their responses to the question of which set of tags is "more fair," given that the auto-tagger would be used in the context of a dating site (RQ1). Following that, we explore their textual explanations for the answers they provided, using the 10 dimensions of fairness discovered in the thematic analysis (RQ2). In particular, we examine the frequency with which these dimensions are used to explain fairness, as well as how they differ by participant gender, as well as by the characteristics of the target image (RQ3).

4.1 Which is more fair?

213 participants (67%) indicated that the human-generated descriptions were "more fair." In contrast, 99 (31%) indicated that Clarifai's

auto-tagger provided a fairer description, while eight participants couldn't tell. A Chi-square test of independence showed no relationship between participant gender and their response.

Table 4 details the proportion of participants who indicated that the human-generated tags were more fair than those generated by Clarifai, broken down by target image. It can be immediately noted that for the images of attractive white individuals (WF-233, WM-004) less than half of the participants indicated that human-generated tags were more fair. Table 4 also presents a logistic regression model (logit model) in which the image code is used, to predict the event that the human-generated tags are perceived as being more fair. We use the following conventions to report statistical significance: *** $p < .001$, ** $p < .01$, * $p < .05$. We also report the odds ratio as a measure of the effect size. The model confirms that for the images of attractive, white individuals (WF-233, WM-004), human-generated tags are less likely to be seen as more fair, as compared to those produced by Clarifai's algorithm.

Table 4: Logit model to predict the event that human-generated tags are perceived as being more fair.

	Human more fair	Estimate	Z	Odds ratio
Intercept (BF-231)	.78	1.237	3.266^{**}	3.44
WF-036	.93	1.276	1.797	3.58
BF-233	.70	-3.895	-0.760	0.68
WF-233	.48	-1.337	-2.708^{**}	0.263
BM-009	.65	-0.6177	-1.227	0.54
WM-022	.75	-1.382	-0.263	0.87
BM-234	.78	-4.498	0.000	1.00
WM-004	.28	-2.206	-4.256^{***}	0.110

4.2 Explaining fairness

Table 5 presents the total number of explanations in which each of the 10 dimensions of fairness is mentioned. In addition, it details the proportion of explanations, by participant gender, using a z-test to flag statistically significant differences by gender. As can be seen, there is only one difference, with men being more likely than women to mention the depicted person's physical characteristics in their explanations of fairness. In Table 6, the pairwise co-occurrence of the dimensions in explanations is examined. As observed, many explanations discuss the accuracy of the descriptions with respect to the depicted person's physical characteristics. Likewise, many invoke the objectivity/subjectivity (i.e., abstract/concrete) characteristics and the accuracy of the tags.

Finally, we explore the possibility that the target person's characteristics might correlate to the dimensions of fairness used in an explanation. Table 7 examines the proportion of explanations referring to each dimension, broken out by the eight target images being described. As observed, there are no striking differences in the manner that participants explain fairness, as a function of the characteristics of the person being described by the taggers. Although there is some variance between images (e.g., participants viewing the image WM-022 did not discuss the issue of accuracy in their explanations as often as those viewing the other images), there appear to be no systematic differences by the depicted person's physical attractiveness, gender or race. In other words, regardless of the target image, participants often evaluated the taggers on accuracy, physical characteristics, and objectivity/subjectivity, in the dating site context.

Table 5: Use of fairness dimensions by participant gender.

	Men (n=160)	Women (n=160)	Z
Accuracy (n=214)	0.71	0.63	1.52
Physical (n=147)	0.51	0.40	1.98^{*}
Obj./Sub. (n=138)	0.47	0.39	1.44
Understanding (n=50)	0.17	0.14	0
Political Corr. (n=45)	0.13	0.16	0
Race (n=39)	0.13	0.12	0.74
Biases (n=24)	0.09	0.06	1.01
Racist (n=9)	0.01	0.04	-1.71
Gender (n=3)	0.01	0.006	0.40
Age (n=0)			

4.3 When is an algorithm more fair?

As mentioned, for six of the eight images, participants generally found the human-generated descriptions to be more fair than those produced by Clarifai. However, for two images, the reverse was true. Table 8 examines the reasons why participants might judge the algorithm to be more fair. In particular, the proportion of explanations referencing each dimension is broken out by participants' fairness answer (i.e., Human/Clarifai).

Although it is not surprising that "Accuracy" is the most frequently mentioned theme in the explanations (214 or 67% of all explanations mention accuracy), it is interesting that it was more frequently used when participants explained a choice that the algorithm's tags were more fair. In contrast, participants who reported human-generated tags to be more fair, were more likely to discuss the balance between objective and subjective attributes in the tags.

Another significant difference concerned the dimensions of Context (e.g., mentioning the use of the tags in the dating site scenario) as well as Political Correctness. These attributes were used more frequently when participants explained why they perceived Clarifai tags as being more fair. Finally, participants discussed Racism more in explanations in which Clarifai was seen as more fair. As previously illustrated in Figure 2, Clarifai's tags are generally more interpretive (i.e., abstract) as compared to human-generated tags; some participants found this to be desirable in the context of a dating site while others did not. Political correctness was also invoked in reference to human-generated tags such as "big lips," "bigger ears" or "shiny skin," in contrast to Clarifai's more conservative word choices.

5 DISCUSSION

Previous work has demonstrated that implicitly extracting information about users in reciprocal recommender systems, can help in producing more accurate recommendations [19]. However, when

Table 6: Number of explanations on fairness judgments in which two dimensions co-occur.

	Accuracy	Physical	Obj./Sub.	Understanding	Political Corr.	Race	Biases	Racist
Physical	102							
Obj./Sub.	79	63						
Understanding	37	24	12					
Political Corr.	15	20	17	4				
Race	21	25	9	2	3			
Biases	10	12	10	0	5	5		
Racist	4	3	3	0	4	5	3	
Gender	3	0	0	0	0	2	1	1

Table 7: Proportion of explanations mentioning each dimension, by image.

	Average				Attractive			
	BF-231	WF-036	BM-009	WM-022	BF-233	WF-233	BM-234	WM-004
Accuracy	0.65	0.83	0.63	0.38	0.60	0.63	0.78	0.88
Physical	0.48	0.40	0.38	0.50	0.45	0.58	0.50	0.40
Obj./Sub.	0.60	0.63	0.35	0.45	0.43	0.43	0.33	0.25
Understanding	0.08	0.03	0.30	0.23	0.13	0.08	0.38	0.05
Political Corr.	0.15	0.05	0.23	0.13	0.08	0.28	0.05	0.18
Race	0.20	0.13	0.13	0.08	0.23	0.08	0.15	0
Biases	0.10	0.03	0.13	0.10	0.08	0.05	0.08	0.05
Racist	0.05	0	0.13	0	0	0	0.05	0
Gender	0	0	0.03	0	0.03	0	0.03	0

Table 8: Proportion of explanations referring to each dimension, by answer.

	Human fair (n=213)	Clarifai fair (n=99)	Z
Accuracy (n=214)	0.52	0.74	-3.68***
Physical (n=147)	0.46	0.49	-0.49
Obj./Sub. (n=138)	0.52	0.25	4.48***
Understanding (n=50)	0.18	0.11	1.58
Political Corr. (n=45)	0.07	0.30	-5.40***
Race (n=39)	0.10	0.18	-1.99*
Biases (n=24)	0.07	0.08	-0.32
Racist (n=9)	0.01	0.06	-2.59**
Gender (n=3)	0.005	0.02	-1.25
Age (n=0)			

we employ automated image tagging for user modeling, especially in reciprocal systems, we need to understand the “fairness" limitations as well as the advantages and the impact these will have on the user we model and his/her in-system relationships. In this section, we relate our findings to the research questions posed. In addition, we discuss avenues for future research, as well as the limitations of our current approach.

5.1 Which description is “more fair"?

Generally speaking, human-generated tags over the algorithm-generated tags were judged as being more fair. The exceptions were the two images of white, attractive individuals, for which Clarifai's tags were seen as more fair. In other words, neither the algorithm- nor the human-generated tags were seen as definitively more fair across all images.

5.2 Key factors in judging fairness

When asked to explain their preference between the descriptions with regard to fairness, the participants discussed 10 reoccurring themes, which users consider to make a difference in terms of a tagger's fair handling of people images. The dimensions most often discussed were Accuracy, Physical Characteristics, and Objectivity/Subjectivity. A key reason Accuracy came up so often is probably that inaccurate tags would exclude users from recommendations in the dating site, reducing either the number of potential matches or the quality of the recommendations based on this tag. The Accuracy dimension also contains responses that refer to the specificity of the tags; some users contrasted the "eye" tag (from the algorithm) with the "blue eyes" tag (from the human). While the general tags may yield more recommendations, the specific tags may yield more precise (and successful) recommendations.

Participants mentioning physical characteristics were often using them to discuss the accuracy (69% co-occur with Accuracy) or the verifiability (i.e. Objectivity, of which 46% co-occur with Physical Characteristics) of the descriptions. Some objective attributes (“long hair," “dark skin") could be useful for implicitly inferring a user's aesthetic preferences in the dating site context.

The third most common dimension, Objectivity/Subjectivity, refers to responses which discuss the objectivity/subjectivity of the descriptions (e.g. "eye," "hair," "serious," "cute"). As seen in Figure 2, Clarifai's tags for the images we used were more likely to include abstract tags while the human-generated tags were more likely to include concrete tags. In fact, one participant noted that *"[the human-generated description] is more data oriented"*. Objectivity/Subjectivity, or whether the tags pass judgment on the individual in the image, has major implications for reciprocal recommendations. A common concern for our participants, the judgment of a machine may not hold true for all users who will unknowingly be affected by the decisions using these subjective tags as a basis for their recommendations. It is important to note however that while some participants felt that including subjective tags was unfair, others felt it was actually more fair, especially in the context of dating. Thus, potential recommendations, and users' perception of the fairness of the system, will all be affected by the use (or absence) of subjective tags that will be used in the user models. In the event that these tags are revealed to the users regarding their own images (e.g. scrutable user modeling), a subjective tag (or the lack of a specific one) may impact the user's self worth [23].

Similarly, the dimensions of Understanding and Political Correctness, which refer to how the users perceive the tags (and by extension, the person in the image) as well as Biases and Racism, which refer to the possibility of systematic and/or extreme differences in how various social groups are represented by the tags, were also brought up in the participants' descriptions. The fact that such aspects are noted suggests that participants/users often pay attention to how these tags may represent themselves or others in the system. Tags that do not meet the users' standards of fairness regarding these dimensions may impact their psychological well-being [21]. Furthermore, it could also affect the quality of the recommendations since certain tags may rarely/never be used for certain social groups (e.g., attractiveness features may be attached more often to white women as compared to other groups).

5.3 User - and image - attributes

As discussed earlier, the algorithm-generated description was perceived as being "more fair" when the depicted individual was both white and attractive. Given that our experiment involved only two such images, we cannot definitively say which characteristics trigger this difference in perceived fairness. However, it is clear that we cannot assume that algorithmic taggers - or human taggers - will treat people fairly across social groups. In the context of dating, this may affect the quality of the user modeling since the treatment of certain people's images will be significantly different.

Upon examining the dimensions of fairness mentioned with respect to the participant's gender, we can see that men were slightly more likely to discuss physical characteristics than women. This implies that the perception of fairness may change depending on the characteristics of the person creating or judging the image descriptions, as well as the characteristics of the person depicted in the image.

5.4 Limitations and future work

As in all empirical studies, the current work has its limitations, which should be considered when interpreting the results. First, we used "organic" tags from the algorithm and crowdworkers. Future work could manipulate the balance of concrete/abstract tags, only changing up the images of the target person, to establish definitively the effect of the depicted person's attributes on fairness judgments. Another limitation that participants viewed only one image and the two descriptions; if they had viewed two images (e.g., compare an attractive vs. less attractive individual), they might provide richer explanations of fairness. In addition, in analyzing the explanations, we coded only for the presence of the dimensions, and did not record whether they were mentioned in a positive light (i.e., are positively/negatively associated with fairness). Finally, participants were not told that one set of tags was human-generated and one was generated by an algorithm. This information might influence their judgments, and could be investigated in future work.

6 CONCLUSION

Since proprietary image tagging services, such as Clarifai, are not transparent about the list of potential outputs or the process of assigning the outputs/tags to different images, it is important to make sure the service used will treat each social group fairly. In the case of implicit inferences for reciprocal recommendations, specifically for dating, tags produced by algorithms may not always be perceived as fair; perhaps this technology is not yet at a point where it is preferable to the user tagging their own images or indicating their preferences explicitly. Further work may look into whether user-generated tags or implicitly-inferred/algorithm-generated tags are more acceptable for specific tasks or recommendations within reciprocal systems.

7 ACKNOWLEDGMENTS

This project is partially funded by the European Union's Horizon 2020 research and innovation programme under grant agreements No. 739578 (RISE), 810105 (CyCAT) and the Government of the Republic of Cyprus (RISE).

REFERENCES

[1] Oscar Alvarado and Annika Waern. 2018. Towards Algorithmic Experience: Initial Efforts for Social Media Contexts. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 286.

[2] Arthur Aron, Gary W Lewandowski Jr, Debra Mashek, and Elaine N Aron. 2013. The self-expansion model of motivation and cognition in close relationships. *The Oxford handbook of close relationships* (2013), 90–115.

[3] Pinar Barlas, Kyriakos Kyriakou, Styliani Kleanthous, and Jahna Otterbacher. 2019. Social B(eye)as: Human and Machine Descriptions of People Images. In *Proceedings of the 13th Annual Conference on Web and Social Media (ICWSM '19)*. AAAI.

[4] Reuben Binns. 2018. Fairness in Machine Learning: Lessons from Political Philosophy. In *Conference on Fairness, Accountability and Transparency*. 149–159.

[5] Rebecca J Brand, Abigail Bonatsos, Rebecca D'Orazio, and Hilary DeShong. 2012. What is beautiful is good, even online: Correlations between photo attractiveness and text attractiveness in men's online dating profiles. *Computers in Human Behavior* 28, 1 (2012), 166–170.

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

[7] Trudy Hui Hui Chua and Leanne Chang. 2016. Follow me and like my beautiful selfies: Singapore teenage girls' engagement in self-presentation and peer comparison on social media. *Computers in Human Behavior* 55 (2016), 190–197.

[8] Michael A DeVito, Jeremy Birnholtz, Jeffery T Hancock, Megan French, and Sunny Liu. 2018. How People Form Folk Theories of Social Media Feeds and What It Means for How We Study Self-Presentation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 120.

[9] Nicholas Diakopoulos. 2016. Accountability in algorithmic decision making. *Commun. ACM* 59, 2 (2016), 56–62.

[10] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. I always assumed that I wasn't really that close to [her]: Reasoning about Invisible Algorithms in News Feeds. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. ACM, 153–162.

[11] Tarleton Gillespie. 2014. The relevance of algorithms. *Media technologies: Essays on communication, materiality, and society* 167 (2014).

[12] Matthew Kay, Cynthia Matuszek, and Sean A Munson. 2015. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 3819–3828.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[14] Debbie S Ma, Joshua Correll, and Bernd Wittenbrink. 2015. The Chicago face database: A free stimulus set of faces and norming data. *Behavior research methods* 47, 4 (2015), 1122–1135.

[15] Evelyn P Meier and James Gray. 2014. Facebook photo activity associated with body image disturbance in adolescent girls. *Cyberpsychology, Behavior, and Social Networking* 17, 4 (2014), 199–206.

[16] Cathy O'Neil. 2017. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

[17] Luiz Pizzato, Thomas Chung, Tomek Rej, Irena Koprinska, Kalina Yacef, and Judy Kay. September 2010. Learning user preference in online dating. http://www.ke.tu-darmstadt.de/events/PL-10/papers/8-Pizzato.pdf In: Hüllermeier, E., Fürnkranz, J. (eds.), Proceedings of the Preference Learning (PL-10) Tutorial and Workshop, European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD). http://www.ke.tu-darmstadt.de/events/ PL-10/papers/8-Pizzato.pdf.

[18] Luiz Pizzato, Tomasz Rej, Joshua Akehurst, Irena Koprinska, Kalina Yacef, and Judy Kay. 2013. Recommending people to people: the nature of reciprocal recommenders with a case study in online dating. *User Modeling and User-Adapted Interaction* 23, 5 (01 Nov 2013), 447–488. https://doi.org/10.1007/s11257-012-9125-0

[19] Luiz Pizzato, Tomek Rej, Thomas Chung, Irena Koprinska, Kalina Yacef, and Judy Kay. 2010. Reciprocal Recommender System for Online Dating. In *Proceedings of the Fourth ACM Conference on Recommender Systems (RecSys '10)*. ACM, New York, NY, USA, 353–354. https://doi.org/10.1145/1864708.1864787

[20] Lauren Rhue. 2018. Racial Influence on Automated Perceptions of Emotions. *Available at SSRN 3281765* (2018).

[21] Carol D Ryff. 1989. Happiness is everything, or is it? Explorations on the meaning of psychological well-being. *Journal of personality and social psychology* 57, 6 (1989), 1069.

[22] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry* (2014), 1–23.

[23] Michael A Stefanone, Derek Lackaff, and Devan Rosen. 2011. Contingencies of self-worth and social-networking-site behavior. *Cyberpsychology, Behavior, and Social Networking* 14, 1-2 (2011), 41–49.

[24] Michele Wilson. 2017. Algorithms (and the) everyday. *Information, Communication & Society* 20, 1 (2017), 137–150.

[25] Peng Xia, Shuangfei Zhai, Benyuan Liu, Yizhou Sun, and Cindy Chen. 2016. Design of reciprocal recommendation systems for online dating. *Social Network Analysis and Mining* 6, 1 (10 Jun 2016), 32. https://doi.org/10.1007/s13278-016-0340-2

[26] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 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*.

[27] Yong Zheng, Tanaya Dave, Neha Mishra, and Harshit Kumar. 2018. Fairness In Reciprocal Recommendations: A Speed-Dating Study. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization (UMAP '18)*. ACM, New York, NY, USA, 29–34. https://doi.org/10.1145/3213586.3226207