Emotion-based Stereotypes in Image Analysis Services

Kyriakos Kyriakou*
k.kyriakou@rise.org.cy
Research Centre on Interactive Media, Smart Systems & Emerging Technologies (RISE Ltd.)
Nicosia, Cyprus

Styliani Kleanthous
styliani.kleanthous@ouc.ac.cy
Cyprus Center for Algorithmic Transparency, Open University of Cyprus
Nicosia, Cyprus

Jahna Otterbacher
jahna.otterbacher@ouc.ac.cy
Cyprus Center for Algorithmic Transparency, Open University of Cyprus
Nicosia, Cyprus

George A. Papadopoulos
george@cs.ucy.ac.cy
Department of Computer Science, University of Cyprus
Nicosia, Cyprus

ABSTRACT

Vision-based cognitive services (CogS) have become crucial in a wide range of applications, from real-time security and social networks to smartphone applications. Many services focus on analyzing people images. When it comes to facial analysis, these services can be misleading or even inaccurate, raising ethical concerns such as the amplification of social stereotypes. We analyzed popular Image Tagging CogS that infer emotion from a person's face, considering whether they perpetuate racial and gender stereotypes concerning emotion. By comparing both CogS and Human-generated descriptions on a set of controlled images, we highlight the need for transparency and fairness in CogS. In particular, we document evidence that CogS may actually be more likely than crowdworkers to perpetuate the stereotype of the "angry black man" and often attribute black race individuals with "emotions of hostility".

CCS CONCEPTS

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

KEYWORDS

computer vision, face analysis, emotion tagging

ACM Reference Format:
Kyriakos Kyriakou, Styliani Kleanthous, Jahna Otterbacher, and George A. Papadopoulos. 2020. Emotion-based Stereotypes in Image Analysis Services. In *Adjunct Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '20 Adjunct), July 14–17, 2020, Genoa, Italy.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3386392.3399567

1 INTRODUCTION

Recent years have seen tremendous growth in Artificial Intelligence (AI), with advances in research as well as the market sector of real-life applications. Given the high demand for AI components and functionalities, many tech companies and start-ups are aligning themselves with the AI industry. According to an article published by Forbes,[1] nearly half of all "AI start-ups" are cashing in on the hype as statistics show that they attract 15% to 50% more funding than other technology firms.

Given the excitement, industry newcomers focus on incorporating AI technologies in their end-user solutions, aiming to solve real-world complex problems and provide innovative features. Many investors, industry leaders and end-users show their preference for solutions that fuse AI techniques.[2] Cognitive Services (CogS) is one of the latest industry trends. Trusted providers offer paid access to state-of-the-art components (e.g., vision, language, speech, search), making it easy for developers to incorporate AI into their products/apps. Microsoft even describes "democratizing AI." [3]

Indeed, over the last few years, many user-centric products have arisen using AI technologies, which rely on understanding user emotion. FeelyBot[4] is a characteristic example, which aims to improve the day of a user through positive interactions via the Jibo[5] platform, helping the individuals treat behavioral and emotional disorders such as depression and social anxiety from which they might suffer. To learn how the user is feeling based on visual data (i.e., images collected during interaction), it combines Microsoft Emotion API[6] to analyze the sentimental elements and Clarifai[7] to get an overall description of an image.

[*] This research constitutes a part of the first author's MSc dissertation at the Department of Computer Science, in University of Cyprus.

UMAP '20 Adjunct, July 14–17, 2020, Genoa, Italy
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7950-2/20/07…$15.00
https://doi.org/10.1145/3386392.3399567

[1] Nearly Half Of All 'AI Startups' Are Cashing In On Hype (March 4, 2019). https://www.forbes.com/sites/parmyolson/2019/03/04/nearly-half-of-all-ai-startups-are-cashing-in-on-hype/

[2] The AI boom is happening all over the world, and it's accelerating quickly (December 12, 2018). https://www.theverge.com/2018/12/12/18136929/artificial-intelligence-ai-index-report-2018-machine-learning-global-progress-research

[3] Democratizing AI For every person and every organization (September 26, 2016). https://news.microsoft.com/features/democratizing-ai/

[4] Clarifai featured hack: Improve your emotional health with FeelyBot (January 5, 2017). https://www.clarifai.com/blog/clarifai-featured-hack-improve-your-emotional-health-with-feelybot

[5] A social robot for the home who looks, listens and learns. https://www.jibo.com)

[6] https://azure.microsoft.com/en-us/services/cognitive-services/face/

[7] https://www.clarifai.com/

Another example is Woo,[8] a dating application targeting people of the Indian diaspora globally. With a "women-first approach," it focuses on creating meaningful relationships and conversations. Woo uses Amazon Rekognition[9] to derive rich image metadata from profile pictures, to aid the automated curation of new user-profiles and photos according to Woo quality guidelines. The quality includes a number of parameters such as the number of faces, size of faces, and estimated age range of the depicted person.

1.1 Ethical concerns

CogS have enabled developers to encapsulate computer vision capabilities into their creations without having to have in-house machine learning experts. However, the research community has demonstrated many ethical issues surrounding computer vision. Goldenfein discusses its potential to profile people [13]. He explains the effect of "Computational Empiricism" in these systems, explaining the need to consider the trustworthiness of the extracted hidden meanings from images that often elude human judges. As he states, *"computer vision systems do not see - they measure"*; indeed, many have scrutinized the predictions of those systems, which can inadequately represent the people depicted in images.

For instance, Kyriakou and colleagues examined the issue of fairness in the descriptions of image tagging CogS [2, 18]. Although they examined how image tagging CogS treat different groups of people, their analyses were limited to the general tagging models of various CogS; they did not examine the behaviours of those specifically designed to process images of people. In the current work, we focus on CogS meant for analyzing very human aspects of the faces depicted in people images. Specifically, we consider Emotion Analysis Services (EAS), an area ripe for analysis given the recent developments in the psychology literature surrounding Emotion Theory, as will be presented.

The UMAP community has been concerned with emotion extraction and usage in the context of recommender and personalization systems. Polignano et al. [21] highlight the need for crafting efficient and personalized strategies to increase customer loyalty. They emphasize the importance of the social and psychological aspects and their ability to aid or replace human decision-making tasks. Tkalčič et al. [24] surveyed the work on the usage of personality and emotions in recommender systems. They noted that personality and emotions account for a good deal of variance in human decision making. Biel et al's [4] work is based on the fact that the human face is an important source of information in interpersonal impressions. They address the problem of predicting vloggers' personality impressions from automatically extracted facial expressions of emotion. Zheng [25] agrees on the effectiveness of users' emotions as contextual information in recommender systems. The author proposes the incorporation of emotional reactions as regularization terms in the context-aware matrix factorization approach, and further explores its effects on the performance of recommendations. Both works [24, 25] emphasize that because emotions can change quickly, it becomes challenging to model and capture them.

Since algorithms are *socio-technical artifacts*, they are often influential in shaping outcomes by judging people, denying or permitting them to access opportunities through the applications in which they are used. Beer emphasizes that algorithms are able to influence and convince people, and describes how they became widely trusted for their precision and objectivity [3]. As a consequence, they can adhere or form truths that can amplify or enrich the social perception of society on a global scale. That is why it is of particular concern that many CogS focus specifically on providing automated analyses on images depicting a human face.

Andalibi et al. [1] shed light on what harms emotion recognition technologies must take into consideration. They surveyed users' attitudes towards emotion recognition technologies citing the need for the socially responsible use and treatment of data in algorithmic decision-making that impacts personal lives. The authors associate their findings with two kinds of perceived risks on emotion recognition: *individual* and *societal* risks. Their participants were particularly concerned about their emotional data, because they felt that emotions could be easily manipulated to impact behavior. In addition, they characterised emotion recognition as invasive, especially in terms of privacy. Further, the authors found that the possibility for unfair and inaccurate interpretations and lack of control over one's digitally curated image by emotion recognition can impact people during and well beyond their lifetime. As the impact of emotion recognition can go beyond the individual, having political and social influence, concerns about the lack of responsibility and regulation with algorithms (accountability matters) were also raised by the participants during the survey.

Today, there is a wide variety of Image Tagging Services (ITAs) offered as CogS that can be used to infer subjective metadata from people images such as traits, emotions and more. Amazon Rekognition Image, Clarifai, Google Cloud Vision,[10] Imagga Auto-tagging,[11] Microsoft Computer Vision and IBM Watson Visual Recognition[12] are only a few of the best known ITAs. Currently, we focus on the Emotion Analysis Services (EAS) of the above tools, which label an input person image with the predicted set of emotions, based on the depicted person's facial expression. The services can be characterized as black boxes; thus, it is necessary to gauge whether or not they treat the input images in an ethical manner.

1.2 The problem

As will be detailed in Section 2, emotion analysis is challenging even for experts on Emotion Theory. For instance, it is difficult to define distinct emotions in a universal manner. This is because of many factors, including the age, gender and cultural background - both, the person making the judgment and the person being judged - influence the perception of emotion. Additional influences include one's visual external characteristics as well as the environmental background or context. These biases may affect emotion perception of EAS as they are trained on human-generated data. Despite the difficulties in human emotion detection, EAS function in a manner that makes these judgments seem easy and objective. Tables 1, 2 and 3 show examples of how EAS interpret the individuals in Figure 1.

[8] https://getwooapp.com/
[9] https://aws.amazon.com/rekognition/
[10] https://cloud.google.com/products/ai/
[11] https://imagga.com/solutions/auto-tagging.html
[12] https://www.ibm.com/watson/services/visual-recognition/

As can be seen, Microsoft outputs probabilities on eight emotions, while Amazon offers a confidence score (from 0 to 100) on the seven emotions. In contrast, Google returns a label describing the likelihood of each emotion characterizing the individual.

Despite that these faces have a neutral expression, EAS often cannot recognize the depicted individuals' neutral emotion. Table 1 shows that Microsoft marked BF-203 (Black Female) with the highest scores on sadness and neutral emotions. Table 3 shows that Google marked BM-213 (Black Male) with a higher score for joy. Furthermore, Table 2 shows that Amazon marked WF-241 (White Female) with a relatively high score for sad and calm, while also it scored WM-220 (White Male) high on angry and calm.

1.3 Goals of the current work

Previous work suggested that image processing algorithms do not treat depicted people fairly. Rhue examined the output emotion from services of Microsoft Face API and Face++[13] using a dataset consisting of black and white basketball players [22]. She found evidence that Face++ interprets black players as angrier than whites, even when controlling their degree of smiling, which was found to be unrelated. In addition, Microsoft infers *contempt* instead of *anger* and interprets black players as more contemptuous when their facial expression is ambiguous. In other words, EAS might exhibit racial bias when inferring emotion. A particular concern is that EAS may be prone to propagating or amplifying - implicitly or explicitly - specific emotion stereotypes that are prevalent in society, when used in applications that are widely deployed.

Therefore, we present an audit of three popular EAS: Amazon Rekognition Image (hereon: Amazon), Google Vision (hereon: Google) and Microsoft Computer Vision (hereon: Microsoft). Given the challenges, the main research question driving our audit, is: "to what extent do EAS perpetuate gender- and race-based stereotypes concerning emotion?"

Table 1: Microsoft Emotion Prediction examples.

Emotion	BF-203	BM-213	WF-241	WM-220
Sadness	0.568	0.003	0.064	0
Neutral	0.428	0.988	0.935	0.988
Contempt	0.001	0.003	0	0.008
Disgust	0.001	0	0	0
Anger	0	0	0	0.004
Surprise	0.001	0	0	0
Fear	0.001	0	0	0
Happiness	0	0	0.006	0

2 BACKGROUND

Here, we review literature on emotion theory and perception, to develop questions to be answered through our audit of the EAS.

2.1 Studying emotion

Ever since Darwin's 19th-century proposals on the nature of emotion, psychologists have been trying to agree on the basic set of emotions found across cultures [6]. To date, the literature is fragmented, with no clear consensus. Many researchers of emotion follow the 1992 suggestions of Ekman and colleagues [8], who found evidence in support of six emotions: happiness, surprise, fear, sadness, anger, and disgust (including contempt) [9]. The majority of the emotion theory experts use these basic six emotions. However, others categorize emotions in abstract and generic classes such as positive, negative and neutral emotion [14]. We consider the same set of basic six emotions in our study. In addition to those six, we added a Neutral option, for cases where the emotions cannot be categorized under a specific emotion from the basic six.

Table 2: Amazon Emotion Prediction examples.

Emotion	BF-203	BM-213	WF-241	WM-220
Happy	5.15	2.80	0.46	5.15
Sad	23.90	5.86	62.22	1.13
Angry	16.50	15.95	1.95	63.00
Confused	7.51	23.85	1.78	8.03
Disgusted	26.84	16.52	0.49	2.11
Surprised	4.38	7.33	0.48	2.63
Calm	15.72	27.69	32.62	17.64

2.2 Social stereotyping and potential harm

For many years now, Emotion Theory experts have studied the social stereotypes surrounding emotion perception. For instance, many researchers investigate the social attributes (e.g., gender, age, race) that can influence the stimuli of emotion perception and as a consequence, the prejudices against the perceived person.

Gender influences on the Perception of Emotion. Lindeberg et al. investigated whether gender and attractiveness moderate emotion perception [19]. They found that happy faces are categorized as "happy" faster than angry faces are recognized as "angry," a phenomenon known as *"the happy face advantage"*. However, when they analyzed the female faces separately, a happy face advantage was found on the attractive females, but not for the unattractive females. Additionally, when both genders were categorized together, the evidence supported their attractiveness hypothesis, while also underscoring the moderating effect of gender.

Furthermore, angry expressions on male faces were shown to be recognized faster, as compared to angry faces on females. Generally, male faces of non-white races, when expressing a negative emotion, were shown to be categorized faster than happy expressions. Finally, sadness was more strongly associated with females and anger with males of any race, but more specifically with the black race faces. This supports the notion that social attributes such as the *race*, *gender* and *attractiveness* of a target person can influence the process of emotion perception.

In another study, Fabes et al. extended the notion of the stereotypes of emotionality by examining the gender and age differences in emotion perception [10]. More specifically, they observed that children perceived as expressing anger are more likely to be boys, while fear and sadness are more often attributed to girls, leading to the stereotype that *"boys get angry, girls get sad"*.

[13] https://www.faceplusplus.com

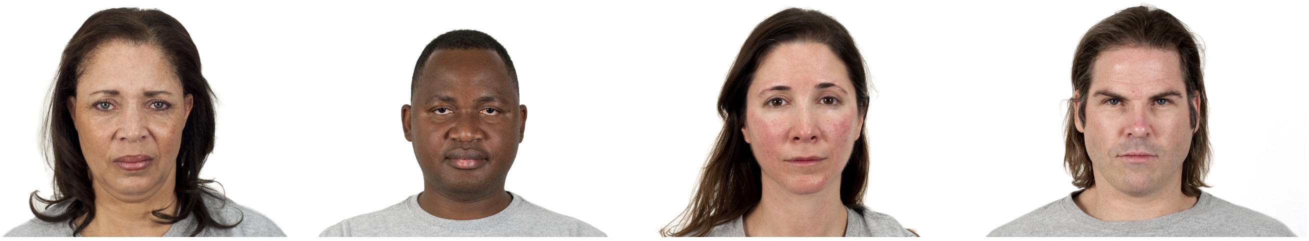


Figure 1: CFD images of black (BF-203, BM-213) and white (WF-241,WM-220) individuals with Neutral facial expressions.

Table 3: Google Emotion Prediction examples.

Emotion	BF-203	BM-213	WF-241	WM-220
joy_likelihood	VERY_UNLIKELY	LIKELY	VERY_UNLIKELY	VERY_UNLIKELY
sorrow_likelihood	UNLIKELY	VERY_UNLIKELY	VERY_UNLIKELY	VERY_UNLIKELY
anger_likelihood	VERY_UNLIKELY	VERY_UNLIKELY	VERY_UNLIKELY	VERY_UNLIKELY
surprise_likelihood	VERY_UNLIKELY	VERY_UNLIKELY	VERY_UNLIKELY	VERY_UNLIKELY

In summary, the literature provides evidence that emotion attribution is gender-stereotyped. Females are typically associated with emotions like sadness and fear. Males, on the other hand, are more likely to be considered as angrier and more aggressive. There is a prevalent social belief that females are more emotional than males (i.e., the “warm women" and “agentic men" stereotype) [5, 11]. This belief implies that males experience and express less emotion, less frequently, and with more self-control than females do [10].

Race influences on the Perception of Emotion. Research has demonstrated differences in emotion perception when one is judging someone of her own race versus another. In one study [14], Chinese participants attributed Caucasian faces with more positive emotion as compared to individuals of their own race, supporting the stereotype of Caucasians as more emotionally positive as compared to others. In contrast, in the same study, Korean participants perceived neutral Caucasian faces as being negative. In another study, Hugenberg et al. provided evidence of a prevalent cultural stereotype that African Americans are aggressive [15, 19]. They extended this research by hypothesizing that the racial prejudice is strongly associated with hostile emotions and demonstrated that there is a tendency to categorize the emotionally ambiguous faces of African Americans under emotions of hostility (e.g., anger) [16].

Other influences on the Perception of Emotion. Other social attributes and cultural beliefs can influence one's perception of emotion in others. The emotion of Fear, for example, is often miscategorized as surprise and anger, especially outside of the Western Culture [17]. The same authors also found that languages allow different ways of describing emotions. For example, the English language has words to express 30 distinct emotions, but Chinese provides expressions for 52 distinct emotions. Chinese often combine words or describe an emotion in a short phrase rather than English speakers, who describe the emotion using a single word.

In summary, when it comes to emotion recognition on faces across races and gender, there are many factors – beyond just the physical attributes of a face – that may influence the perception of another's emotions. Social stereotypes (e.g., that women are warm while men are agentic; that blacks are aggressive) clearly influence the process of emotion perception [10]. In this paper, we investigate the extent to which the output of EAS, when interpreting faces of people, perpetuate stereotypes surrounding emotion. In particular, we examine their behaviors when perceiving the faces of black and white individuals.

3 METHODOLOGY

We follow a fused audit methodology following Sandvig et al. [23]. We execute an automated Scraping Audit, issuing a sequence of predefined queries and uploading a standardized image dataset of people faces. We also conduct a crowdsourcing study, having workers tag the same images among a common set of emotions.

We chose the Chicago Face Database (CFD)[14] as our primary dataset [20]. CFD consists of images of faces of 597 diverse individuals including four races and two genders. The races include Asian, Black, Latino and White individuals of males and females between the ages of 18 and 40 years old. CFD is a standardized and normalized dataset, created by psychologists; all people are positioned in the same direction looking straight into the camera in front of a white background. They all wear a grey t-shirt and were instructed not to wear makeup or other personal accessories. For every face in the database, a neutral expression is provided. For a subset of individuals, images of three additional expressions for the emotions of anger, fear, and happiness are provided.

We considered three popular CogS that provide a dedicated EAS:

- Amazon Rekognition Image
- Google Cloud Vision
- Microsoft Computer Vision

3.1 Study 1: Extracting Emotions using EAS

We executed a set of queries to the services using their RESTful APIs via HTTP Requests. Using automated python scripts for each occasion, we uploaded all of the CFD images into these services

[14] https://chicagofaces.org

and stored their JSON formatted responses. We then processed the responses by extracting only the data related to our research: the emotion analysis, creating a middle product of Comma Separated Values (CSV) formatted files. The CSV files consist of attributes such as the identity of the target image processed, the depicted emotion marked from the CFD (e.g. N-Neutral, A-Angry, HO-Happy Open Mouth, HC-Happy Closed Mouth and F-Fear) and the predicted emotions that each service provides. Those emotion categories were divided in distinct rows along with their values.

3.2 Study 2: Extracting Emotions using Crowdsourcing

We used the Figure-Eight[15] Platform for our crowdsourcing tasks. We designed a task as shown in Figures 2 and 3. Figure 2 shows the instructions provided prior to the task. First, we asked workers to *"Help us determine the emotion of the depicted individual in the photo we provide. According to psychologists, there are seven basic emotions including: happiness, anger, surprise, sadness, disgust, fear, and neutral"*. Then we provided some steps to clarify the process:

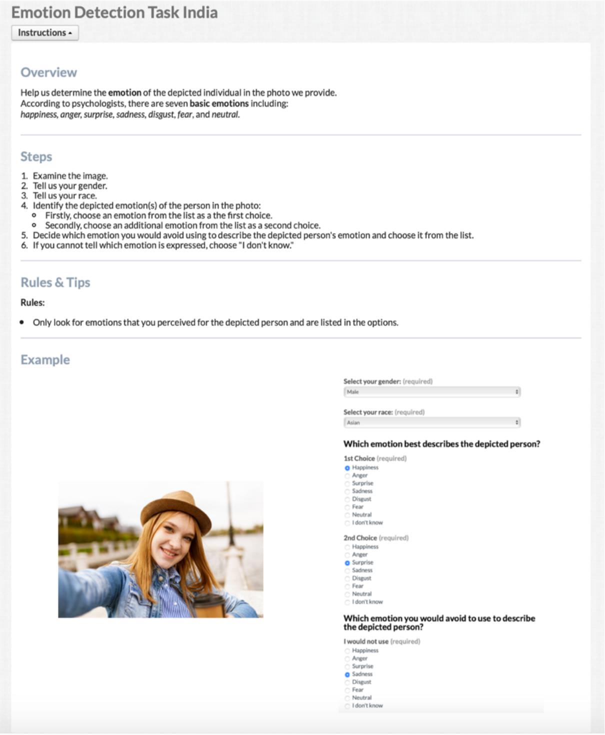
- *Examine the image.*
- *Tell us your gender.*
- *Tell us your race.*
- *Identify the depicted emotion(s) of the person in the photo:*
  - *Firstly, choose an emotion from the list as a the first choice.*
  - *Secondly, choose an additional emotion from the list as a second choice.*
- *Decide which emotion you would avoid using to describe the depicted person's emotion and choose it from the list.*
- *If you cannot tell, choose "I don't know."*

Next to the steps, a rule was attached specifying that *"Only look for emotions that you perceived for the depicted person and are listed in the options."* along with a screenshot of the task as an example.

Task execution. We executed two separate jobs, targeting US and Indian workers. Only the images with a non-neutral expression were passed to the platform as shown in Table 4, in order to use the predefined emotion as a gold standard to observe the deviation of the worker responses (i.e., perceived emotion(s)). Each of the 610 faces with an emotion received three responses from distinct workers. We allowed participants to provide up to 20 judgments (i.e., analyze up to 20 images). Then, we applied a limit criterion on accepting responses with a minimum of 10 seconds to exclude any bots or random responses. We also added a 70% minimum accuracy required for accepting the responses. Both US and India workers were paid 15 cents per task competition. Our task reached 371 India and 321 US participants. As the task took no longer than 60 seconds, this corresponds to an hourly wage of 7.5 USD.

4 FINDINGS

Sections 4.1, 4.2 and 4.3 present the findings based on the Amazon, Microsoft and Google Vision EAS inferences on emotions, respectively, while Section 4.4 presents the findings from the crowdworker study. Within each section, we consider the inferences made on the emotional state of each depicted individual, when he or she is actually depicted with an angry, happy, fearful or neutral expression.



Emotion Detection Task India

Instructions ▴

Overview

Help us determine the **emotion** of the depicted individual in the photo we provide.
According to psychologists, there are seven **basic emotions** including:
happiness, anger, surprise, sadness, disgust, fear, and *neutral*.

Steps

1. Examine the image.
2. Tell us your gender.
3. Tell us your race.
4. Identify the depicted emotion(s) of the person in the photo:
   - Firstly, choose an emotion from the list as a the first choice.
   - Secondly, choose an additional emotion from the list as a second choice.
5. Decide which emotion you would avoid using to describe the depicted person's emotion and choose it from the list.
6. If you cannot tell which emotion is expressed, choose "I don't know."

Rules & Tips

Rules:

- Only look for emotions that you perceived for the depicted person and are listed in the options.

Example

Select your gender: (required)
Male

Select your race: (required)
Asian

Which emotion best describes the depicted person?

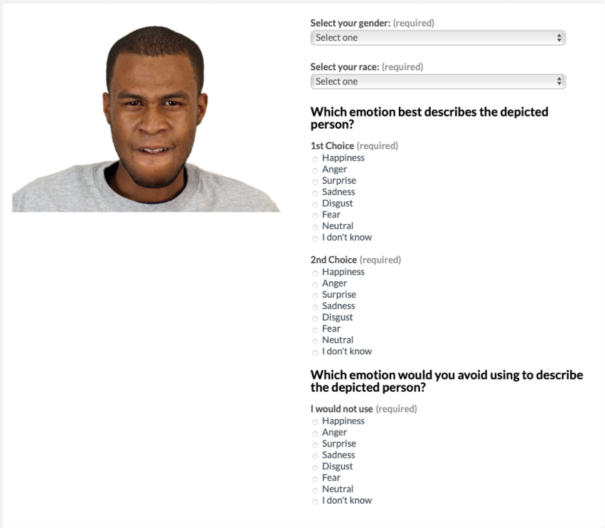
1st Choice (required)
● Happiness
○ Anger
○ Surprise
○ Sadness
○ Disgust
○ Fear
○ Neutral
○ I don't know

2nd Choice (required)
○ Happiness
○ Anger
● Surprise
○ Sadness
○ Disgust
○ Fear
○ Neutral
○ I don't know

Which emotion you would avoid to use to describe the depicted person?

I would not use (required)
○ Happiness
○ Anger
○ Surprise
● Sadness
○ Disgust
○ Fear
○ Neutral
○ I don't know

Figure 2: Crowd-worker Task Instructions



Select your gender: (required)
Select one

Select your race: (required)
Select one

Which emotion best describes the depicted person?

1st Choice (required)
○ Happiness
○ Anger
○ Surprise
○ Sadness
○ Disgust
○ Fear
○ Neutral
○ I don't know

2nd Choice (required)
○ Happiness
○ Anger
○ Surprise
○ Sadness
○ Disgust
○ Fear
○ Neutral
○ I don't know

Which emotion would you avoid using to describe the depicted person?

I would not use (required)
○ Happiness
○ Anger
○ Surprise
○ Sadness
○ Disgust
○ Fear
○ Neutral
○ I don't know

Figure 3: The actual Crowd-worker Task

4.1 Amazon Rekognition

Images with an Angry Expression. A Welch Two Sample t-test was conducted to compare the tagger inference of "angry" between black and white males. To control for the familywise error rate we apply a Bonferroni correction; thus, we use a critical value of 0.0125 rather than 0.05 in evaluating the significance of the four tests

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

Table 4: Number of CFD images with the 4 emotions.

	N	HO	HC	A	F	
BM	104	48	48	47	48	**295**
BW	93	36	32	35	35	**231**
WM	90	35	37	37	37	**236**
WW	93	35	36	35	29	**228**
	380	**154**	**153**	**154**	**149**	990

involving each tagger's outputs. There was a significant difference in the angry scores for black (M=57.679, SD=31.147) and white males (M=39.053, SD=31.685); t (68)=2.746, p=0.008. These results suggest that black males were assigned with a higher score on "anger" than whites, congruent to the stereotype of black people as being angry. However, no significant differences were detected when comparing the angry expressions of black and white women.

Images with a Happy Expression. We treat the CFD images marked as Happy Open Mouth (HO) and Happy Closed Mouth (CO) as one Happy (H) expression. A Welch Two Sample t-test was conducted to compare the "angry" inferences between black and white females on their happy expression. The test indicated that the scores on the angry emotion were significantly greater for black females (M=2.265, SD=2.566) than for white females (M=0.833, SD=0.977), t (129)=5.004, p<.001. Thus, the EAS inferences are in line with the stereotype of black people as being angry. However, no significant differences were detected when comparing the happy expressions of black and white men.

Images with a Fear Expression. The Welch t-test was conducted to compare the "angry" scores between black males and white males on their fear expression, as well as between black females and white females. However, no statistically significant differences were detected. *Images with a Neutral Expression.* No evidence was found of different inferred emotions for black versus white individuals.

4.2 Microsoft Computer Vision

Images with an Angry Expression. A Welch t-test was conducted to compare the angry scores between black and white males on their angry expression. The test indicated that the scores on the angry emotion were significantly greater for black males (M=0.417. SD=0.330) than for white males (M=0.175, SD=0.217), t (59) = 3.628, p<.001. These results suggest the same evidence as those of Amazon; black males were assigned with a higher score for anger than whites, which tends to reinforce the stereotype of black people as being angry. However, no significant differences were detected when comparing the angry expressions of black and white women.

Images with Happy, Fear, Neutral Expressions. No evidence was found of meaningful differences on black versus white faces.

4.3 Google Vision

As Google output is ordinal (i.e., a label expressing the likelihood of the emotion, ranging from VERY_UNLIKELY to VERY_LIKELY), a non-parametric Mann-Whitney Test was used to compare the differences in the inferred emotions between the demographic groups. In parallel to the analyses for the Amazon and Microsoft EAS, we considered the tagger's inferences on images with Angry, Happy, Fear and Neutral facial expressions, to explore the possibility for systematic differences in inferences by the depicted person's demographic group. No statistically significant differences were detected.

4.4 Crowdworker Emotion Perception

Having examined the performance of the EAS on the emotion perception task, we now investigate how well we can expect to fare on the task using a human-in-the-loop approach. First, we consider the extent to which human judges agree on the emotion of a depicted person in an image (i.e., the interjudge agreement on the task). Following that, we investigate the crowdworkers' accuracy on the task, for each of the three ground-truth emotions of interest.

Interjudge agreement.

Table 5 presents the interjudge agreement, based on workers' first response (i.e., first impression as to the depicted person's emotion). There is perfect agreement between judges for around 60% of the images; this is true both of the data provided by workers in India as well as by those located in the US. Given that workers could choose one of eight responses (the six basic emotions, neutral, or "I don't know"), the probability that, for a given image, three judges choose the same response at random is less than 0.02. Thus, we observe high levels of agreement. In addition, full disagreement (i.e., across the three judges, we observe three unique responses) is relatively rare; this happens on fewer than 20% of the images. Given the high levels of agreement, our analyses below consider only the first response provided.

Table 5: Proportion of images on which judges agree.

	IN	US
Three judges agree	.58	.63
Two judges agree	.22	.20
No judges agree	.19	.16

Images with an Angry Expression. We now consider the workers' accuracy on task, when the images being judged depict an angry expression. Table 6 details the accuracy (i.e., percent of images on which "anger" was correctly inferred as the depicted emotion), broken out by the social group of the depicted persons. Accuracy is reported separately for our IN and US crowd data, in order to consider possible cultural differences.

We observe that the US crowdworkers are generally more accurate in interpreting angry expressions, as compared to the workers from India. Perhaps this is because of the familiarity of the faces used by CFD which were individuals from Chicago, US. However, there is no evidence that workers tend to associate *anger* with Black men or women more so than whites.

Images with a Happy Expression. As in the analysis of the algorithmic EAS, we combine the responses to the HO and HC images. Table 6 presents the accuracy of the workers on the images with happy expressions. The workers' accuracy on detecting these expressions is very high. Furthermore, there are no significant differences across the depicted persons' social groups.

Images with a Fearful Expression. In Table 6 we observe that overall, the crowdworkers are not very accurate at spotting a fearful

Table 6: Accuracy on detecting an emotion expression.

	Angry		Happy		Fear	
	IN	US	IN	US	IN	US
BM	44%	67%	89%	92%	22%	23%
BW	48%	70%	91%	93%	18%	17%
WM	50%	67%	91%	84%	27%	18%
WW	52%	71%	98%	95%	27%	23%

expression. However, again, there are no statistically significant differences across the four social groups.

Table 7: Summary of results. "Comparisons" refers to the (predicted emotion / actual emotion) considered.

Stereotype	Comparison(s)	A	G	M	India	US
Black people as angry	BM v. WM (angry/angry)	+		+		
	BW v. WW (angry/happy)	+				
Black people as hostile	BM v. WM (angry/fear)					
Women as happy (warm)	M. v. W (happy/happy)					
Women as sad	M v. W (sad/neutral)					

5 DISCUSSION

The most interesting findings concern the stereotype of black men and women as being more angry or hostile, as compared to whites of the same gender. As discussed in related work [19], the faces of non-white individuals tend to be categorized easier as being angry when they show an angry emotion compared to whites. We confirmed empirically that some EAS (in particular, Amazon and Microsoft) were more likely to correctly infer an angry expression on a black man versus a white man.

Similarly, the Amazon EAS was more likely to incorrectly infer an angry emotion on happy images of black women, versus happy images of white women. This resonates with what was observed in [7, 16] that people tend to categorize ambiguous faces of black individuals under emotions of hostility (e.g., anger).

Furthermore, we found no evidence that Google's EAS treats individuals differently based on their race or gender. Unlike Amazon and Microsoft, Google EAS did not demonstrate any evidence of unwanted race- or gender-based emotion biases.

The Amazon and Microsoft results might lead us to believe that the ground truth data, the EAS are trained on, also reflect such assumptions (e.g., "blacks as angry") and thus, the algorithms learn from those. As proprietary services, we cannot know what datasets are used in training these algorithms. Nonetheless, we analyzed our own dataset of human-generated emotion inferences.

Surprisingly, we found no evidence of the reproduction of these stereotypes by the crowdworkers. It has to be noted that human workers - while not infallible on the emotion recognition task - are quite accurate on this task, and even when they are not accurate, there is a high degree of consensus on their emotion labels. This last finding underscores the difficulty of this task, as was expected from the review of the psychology literature. Table 7 presents a summary of the results of the comparisons on the predicted emotion versus actual emotion and the stereotypes observed per each EAS and crowdworker group. To our knowledge, this study is one of the first in the area of understanding biases in EAS thus, sets the grounds for future work.

6 LIMITATIONS

Every study comes with limitations. Below, we discuss limitations of this study that should be considered when interpreting the results.

Services across providers do not follow the same structure or output format in their emotion analysis. They focus on a different set and number of emotions and are using various metrics for presenting their predictions (e.g., numeric confidence scores vs. ordinal scores). Due to lack of transparency about the models used, further experimentation is needed to understand their behaviour.

Unfortunately, our primary image dataset, CFD, consists of faces of only four races and does not note biracial individuals. Furthermore, for emotional (i.e., non-neutral) expressions, only black and white individuals are depicted.

This study employed crowdworkers in recognizing emotions on images. The task was built to allow 20 judgments per task per participant. This was done mainly for ensuring task completion due to a small pool of participants per location (US and India). We understand that the resulting sample violates the assumption of statistical independence. However, this was an initial study and further work is needed, in a larger scale that will take this limitation into consideration when designing the crowdsourcing task.

Employing crowdsourcing for this study poses further concerns. Unfortunately, we do not have control over how the platform evaluates judgments and of the number of participants that got filtered out due to the accuracy requirements. This might artificially reduce the observed bias in crowdworkers' judgments. In the future, we are planning to mitigate this issue by adding test questions right in the task or (ultimately) use another platform such as Amazon's MTurk to re-execute the crowdsourcing tasks.

7 CONCLUSIONS

We investigated the extent to which emotion-based stereotypes are perpetuated by Emotion Analysis Services (EAS) that are designed to process images of faces. We focused in particular on their treatment of black and white individuals. As noted increasingly by researchers (e.g., [12]), to measure bias and to correct for it, one must first establish a ground truth. To this end, we compared the global behaviors of EAS to that of crowdworkers. We found initial evidence that EAS can perpetuate the stereotypes of the "black angry man," and often attribute black individuals with "emotions of hostility." However, in this study, such evidence was not found in the crowd-generated data. It is clear that developers need tools for identifying, measuring and mitigating social bias in their systems during their development and overall lifetime, examining their inputs and outputs to scrutinize their behaviors for unintended social biases. They have to be aware of such issues to avoid the possible consequences, mitigating the amplification of the biases already prevalent in society.

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] Nazanin Andalibi and Justin Buss. [n. d.]. The Human in Emotion Recognition on Social Media: Attitudes, Outcomes, Risks. ([n. d.]). preprint on Research Gate at https://www.researchgate.net/profile/Nazanin_Andalibi2/publication/338633222_The_Human_in_Emotion_Recognition_on_Social_Media_Attitudes_Outcomes_Risks/links/5e20a28f92851cafc38a83cf/The-Human-in-Emotion-Recognition-on-Social-Media-Attitudes-Outcomes-Risks.pdf.

[2] Pinar Barlas, Kyriakos Kyriakou, Styliani Kleanthous, and Jahna Otterbacher. 2019. What Makes an Image Tagger Fair? Proprietary Auto-tagging and Interpretations on People Images. In *Proceedings of the 27th ACM Conference On User Modelling, Adaptation And Personalization (UMAP '19)*. ACM.

[3] David Beer. 2017. The social power of algorithms.

[4] Joan-Isaac Biel, Lucía Teijeiro-Mosquera, and Daniel Gatica-Perez. 2012. Facetube: predicting personality from facial expressions of emotion in online conversational video. In *Proceedings of the 14th ACM international conference on Multimodal interaction*. 53–56.

[5] Amy JC Cuddy, Susan T Fiske, and Peter Glick. 2008. Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. *Advances in experimental social psychology* 40 (2008), 61–149.

[6] C Darwin. 1872. Facial expression of emotion in man and animals.

[7] Patricia G Devine. 1989. Stereotypes and prejudice: Their automatic and controlled components. *Journal of personality and social psychology* 56, 1 (1989), 5.

[8] Paul Ekman. 1992. Are there basic emotions? (1992).

[9] Paul Ekman, Wallace V Friesen, and Phoebe Ellsworth. 2013. *Emotion in the human face: Guidelines for research and an integration of findings*. Vol. 11. Elsevier.

[10] Richard A Fabes and Carol Lynn Martin. 1991. Gender and age stereotypes of emotionality. *Personality and social psychology bulletin* 17, 5 (1991), 532–540.

[11] Susan T Fiske, Amy JC Cuddy, Peter Glick, and Jun Xu. 2018. A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition (2002). In *Social cognition*. Routledge, 171–222.

[12] Jean Garcia-Gathright, Aaron Springer, and Henriette Cramer. 2018. Assessing and Addressing Algorithmic Bias - But Before We Get There. (sep 2018), 450–454. arXiv:1809.03332 http://arxiv.org/abs/1809.03332https://www.aaai.org/ocs/index.php/SSS/SSS18/paper/view/17542/15470

[13] Jake Goldenfein. 2019. The Profiling Potential of Computer Vision and the Challenge of Computational Empiricism. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 110–119.

[14] Chao S Hu, Qiandong Wang, Tong Han, Ethan Weare, and Genyue Fu. 2017. Differential emotion attribution to neutral faces of own and other races. *Cognition and Emotion* 31, 2 (2017), 360–368.

[15] Kurt Hugenberg and Galen V Bodenhausen. 2003. Facing prejudice: Implicit prejudice and the perception of facial threat. *Psychological Science* 14, 6 (2003), 640–643.

[16] Kurt Hugenberg and Galen V Bodenhausen. 2004. Ambiguity in social categorization: The role of prejudice and facial affect in race categorization. *Psychological Science* 15, 5 (2004), 342–345.

[17] Rachael E Jack, Wei Sun, Ioannis Delis, Oliver GB Garrod, and Philippe G Schyns. 2016. Four not six: Revealing culturally common facial expressions of emotion. *Journal of Experimental Psychology: General* 145, 6 (2016), 708.

[18] Kyriakos Kyriakou, Pınar Barlas, Styliani Kleanthous, and Jahna Otterbacher. 2019. Fairness in Proprietary Image Tagging Algorithms: A Cross-Platform Audit on People Images. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 13. 313–322.

[19] Sofie Lindeberg, Belinda M Craig, and Ottmar V Lipp. 2018. You look pretty happy: Attractiveness moderates emotion perception. *Emotion* (2018).

[20] 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.

[21] Marco Polignano, Pierpaolo Basile, Marco de Gemmis, and Giovanni Semeraro. 2019. Social Tags and Emotions as main Features for the Next Song To Play in Automatic Playlist Continuation. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*. 235–239.

[22] Lauren Rhue. 2018. Racial influence on automated perceptions of emotions. *Available at SSRN 3281765* (2018).

[23] 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* 22 (2014).

[24] Marko Tkalcic, Giovanni Semeraro, and M Gemmis. 2014. Personality and emotions in decision making and recommender systems. RWTH.

[25] Yong Zheng. 2016. Adapt to Emotional Reactions in Context-aware Personalization.. In *EMPIRE@ RecSys*. 1–8.