Proceedings of the Thirteenth International AAAI Conference on Web and Social Media (ICWSM 2019)

Fairness in Proprietary Image Tagging Algorithms: A Cross-Platform Audit on People Images

Kyriakos Kyriakou,[1] **Pınar Barlas,**[1] **Styliani Kleanthous,**[1,2] **Jahna Otterbacher**[1,2]
[1]Research Centre on Interactive Media, Smart Systems and Emerging Technologies (Nicosia, CYPRUS)
[2]Cyprus Center for Algorithmic Transparency, Open University of Cyprus (Latsia, CYPRUS)
{kyriakos093, pin.barlas}@gmail.com, {styliani.kleanthous, jahna.otterbacher}@ouc.ac.cy

Abstract

There are increasing expectations that algorithms should behave in a manner that is socially just. We consider the case of image tagging APIs and their interpretations of people images. Image taggers have become indispensable in our information ecosystem, facilitating new modes of visual communication and sharing. Recently, they have become widely available as Cognitive Services. But while tagging APIs offer developers an inexpensive and convenient means to add functionality to their creations, most are opaque and proprietary. Through a cross-platform comparison of six taggers, we show that behaviors differ significantly. While some offer more interpretation on images, they may exhibit less fairness toward the depicted persons, by misuse of gender-related tags and/or making judgments on a person's physical appearance. We also discuss the difficulties of studying fairness in situations where algorithmic systems cannot be benchmarked against a ground truth.

Introduction

Image analysis algorithms, which infer what is depicted in an input image, represent one of the most widely used applications of computer vision. Since Krizhevsky and colleagues' introduction of the use of deep learning for image classification in the ImageNet Challenge[1] (Krizhevsky, Sutskever, and Hinton 2012), the technology has rapidly improved. More recently, researchers have simplified the overall approach (Simonyan and Zisserman 2014), at the same time developing more efficient ways to manage the necessary computational resources (Szegedy et al. 2015).

Along with the improvements in performance, image analysis has moved beyond more constricted contexts (e.g., analysis of satellite or medical imagery) into consumer applications. We increasingly communicate in a visual manner, and image analysis algorithms enable us to organize or retrieve multimedia content in real time, at the same time filtering out items inappropriate in a given context. These algorithms also influence which content is pushed to consumers; image recognition tools are commonly used in digital asset management, enabling professionals to find and distribute relevant content (e.g., in a digital marketing campaign), and to track its consumption by and popularity with consumers.

Image analysis technology is also transforming commerce, making online and in-store shopping more personalized and convenient.[2] It underpins the application of emerging technologies, such as augmented reality (AR), in this domain. For instance, in a retail setting, image recognition (e.g., detecting a target brand or product in the shopper's physical proximity) could be used to enable the triggering of a personalized digital and/or AR experience. Image recognition and AR APIs specifically tailored to creating such "scan-and-shop" marketing strategies are already available.[3]

Despite the uptake of image analysis in consumer-focused applications, problems can occur, particularly when algorithms process people-related media. High-profile incidents covered in the media have illustrated that image analysis algorithms have the potential to yield socially offensive and discriminatory output, and that the public expects accountability when algorithms behave badly. Most notably was the 2015 Google Photos incident, in which a Black software engineer's photo depicting himself and a friend, was labeled with the tag "gorillas." Google immediately apologized and vowed to find a solution. However, the solution announced in 2018, which involved removing the offending tag from the database, was criticized as an "awkward workaround."[4]

Particularly on social platforms, users carefully craft their self-presentation (Birnholtz et al. 2014). In fact, some findings suggest that social media spaces are projections of one's idealized self (Chua and Chang 2016) and that the practice of uploading "selfies" can contribute to a feeling of self-worthiness (Stefanone, Lackaff, and Rosen 2011). Given incidents such as the above, it is easy to envision how image analysis could adversely affect users' sense of well-being when output tags on images are offensive or otherwise seen as unjust. Therefore, it is crucial to understand how image analysis algorithms treat people-related media, and to develop ways to audit them. This is particularly important because the use of proprietary algorithms by third party de-

[1]http://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://documentation.catchoom.com/

[4]https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/

Figure 1: Example Chicago Face Database (CFD) images of Asian (AF-204), Black (BF-231) & White (WF-200) women.

Tagging API	AF-204	BF-231	WF-200
Amazon Rekognition	human, people, person face, head, portrait dimples	human, people, person Afro, hairstyle hair, face	human, people, person face, portrait, head female, woman
Clarifai API	one, portrait, cute, child, people facial, wear, man, looking face, isolated, funny, adult joy, casual, happiness, pensive adolescent, eye, serious	people, one, portrait, man wear, adult, side, pensive profile, woman, face, isolated child, facial, Afro, casual fashion, athlete, adolescent	woman, portrait, isolated, one, cute casual, people, fashion, eye young, looking, look, pretty young, wear, face, hair, serious adult, friendly, facial
Google Cloud Vision	face, eyebrow, cheek chin, skin, forehead nose, head, jaw, neck	face, forehead, chin, eyebrow cheek, nose, head, jaw, neck, human	face, eyebrow, chin cheek, head, forehead neck, jaw, portrait, ear
Imagga Auto-tagger API	beard, man, face, person, male, portrait handsome, child, guy	Afro, man, face portrait, male handsome, head	beard, portrait, face, person man, attractive, model handsome, male, adult
Microsoft Vision	person, man, necktie, wearing, indoor, shirt, posing, looking suit, camera, glasses, young, photo dress, black, front, standing, neck white, smiling, male, holding, hair	man, person, wearing, looking necktie, standing, shirt, front face, smiling, white, suit posing, hair, holding, neck, young glasses, black, head, hat, red	person, posing, necktie, wearing shirt, young, man, smiling glasses, photo, holding, camera hair, dress, front, standing, black woman, neck, suit, blue, red
Watson Vision	person, skin, light brown color ash grey color	person, woman, female indian red color, coal black color	person, people, face, adult person ash grey color

Table 1: Output tags for example CFD images produced by six image analysis APIs.

velopers is on the rise, through their commercialization, a phenomenon Gartner has called the "Algorithm Economy."[5]

The "democratization" of proprietary algorithms

One industry take on the Algorithm Economy is the creation of Cognitive Services (CogS), which is fueling innovation across sectors. While technology giants have used cognitively inspired algorithms in their products for years, these proprietary algorithms are now being made available to third parties as CogS. Microsoft, one of the leading providers of CogS through its Azure platform, describes them as "democratiz[ing] AI by packaging it into discrete components that are easy for developers to use in their own apps."[6]

While CogS provide convenient and economic solutions to developers looking to add capabilities - like image analysis - to their applications, they are also opaque, at a time when there are rising expectations for developers to be ethical. Professional associations such as the IEEE[7] and the ACM[8] are encouraging developers to take measures to promote transparency in the algorithmic systems they build. Similarly, in the European Union, the General Data Protection Regulation (GDPR), in effect as of May 2018, requires that data controllers be able to explain any automated processes applied to citizens' personal data.[9]

However, CogS exhibit common barriers to transparency (Burrell 2016). They are proprietary and technically complex, such that explaining their behaviors is difficult. Furthermore, the developers' guides for most CogS focus on implementation and do not explain the range of possible outputs for a given algorithmic process (e.g., in image tagging APIs, the full set of tags to describe input images). In other words, the burden of ensuring ethical behavior, such as preventing socially insensitive output, is typically placed entirely on the developer purchasing the use of the CogS.

[5]https://www.gartner.com/smarterwithgartner/the-algorithm-economy-will-start-a-huge-wave-of-innovation/

[6]https://blogs.windows.com/buildingapps/2017/02/13/cognitive-services-apis-vision/

[7]https://ethicsinaction.ieee.org/

[8]https://www.acm.org/binaries/content/assets/public-policy/2017_joint_statement_algorithms.pdf

[9]https://gdpr.algolia.com/gdpr-article-15

Image tagging APIs: interpreting people images

We examine vision-based CogS from six providers: Amazon Rekognition[10], Google Cloud Vision[11], IBM Watson Visual Recognition[12], Microsoft Computer Vision[13], Imagga[14] and Clarifai[15]. Four providers are industry giants that provide a full range of CogS, while Clarifai and Imagga specialize in computer vision services. Across all, we focus on their image tagging services provided via REST APIs.

To study the manner in which these APIs interpret images of people, we use the Chicago Face Database (CFD) (Ma, Correll, and Wittenbrink 2015) as our primary dataset. CFD consists of photographs of men and women from four racial backgrounds, which have been taken in a very controlled manner. All persons are wearing the same grey t-shirt and are positioned in a neutral manner and shown in Figure 1. Through our analysis, we address three research questions:

1. Given that many APIs do not disclose their full set of tags, what are the key differences, in terms of the descriptive tags used, across APIs?
2. Do proprietary image tagging algorithms interpret all people images in a similar manner?
3. How can we approach the notion of fairness in a setting where algorithmic output is open-ended, such that there is a wide range of "correct" responses?

Examining tagger behavior. Figure 1 shows example photos from the CFD. Table 1 details the output tags provided by each API for each photo. Here, several initial observations can be made, with respect to the manner in which the depicted persons are interpreted by the algorithms:

- **Gender inference:** While taggers are general tools and are not specifically designed to recognize gender, many gender-related tags are used, often inaccurately.
- **Judgment tags:** Some tags are subjective in nature and do not logically follow from the content of the input image. In particular, many comment on a person's physical appearance (e.g., attractive, cute, pretty).
- **Abstract inferences:** Similarly, we observe tags that describe one's occupation or role (e.g., athlete) as well as perceived character traits or emotional states (e.g, serious, friendly, pensive). These characteristics are not concrete, based on what can be observed in the photo.

Developers who use these APIs in their creations, may not expect the above behaviors. A natural question to answer, before using these APIs is, do they interpret people - regardless of gender or race - in a fair manner? Do they describe similar aspects or characteristics in their output tags? To this end, we carry out an evaluation of the tags output by the six algorithms for the CFD images, to characterize and compare their behaviors when interpreting men and women of different races. Finally, we relate our findings to the broader conversation in the research community surrounding fairness and transparency in algorithmic systems.

[10]https://aws.amazon.com/rekognition/

[11]https://cloud.google.com/vision/overview/docs/

[12]https://www.ibm.com/watson/services/visual-recognition/

[13]https://azure.microsoft.com/en-us/services/cognitive-services/computer-vision/

[14]https://imagga.com

[15]https://clarifai.com

Background

There has been an explosion of interest in the behaviors of algorithmic systems, and in particular, how to detect and redress their biases. Attention to the topic arguably stems from the growing influence of opaque - usually proprietary - algorithms in our information ecosystem. Algorithms have become "power brokers" that are not always held accountable for their decisions and actions (Diakopoulos 2016).

Algorithms are increasingly delegated everyday tasks and operate largely autonomously, without human intervention (Wilson 2017). In fact, human behavior tends to reinforce the power and autonomy of algorithms; there is a tendency for people to perceive them as objective (Gillespie 2014; O'Neil 2017) while some remain totally unaware of algorithmic interventions in the systems they use (Eslami et al. 2015). Automated content analysis on images is a prime example of an "everyday" task which, as previously mentioned, underlies many other modern applications.

Here, we lay the groundwork for conducting a within- and cross-platform audit of image tagging APIs when interpreting images of people. The study provides insights on these opaque processes, in light of the growing awareness of possible social harms stemming from algorithmic biases. At the same time, we connect our work to the ongoing conversation in the research community surrounding the issue of fairness in algorithmic systems, in contexts like image tagging where it is difficult to define an objective notion of algorithmic bias.

Auditing algorithms: two approaches

Given that image tagging CogS are proprietary, how might we evaluate their behaviors when analyzing images of people? Sandvig and colleagues (Sandvig et al. 2014) suggested auditing algorithms "from the outside" when full transparency (i.e., a code inspection) is not possible. Eslami and colleagues (Eslami et al. 2017) articulated two approaches. In a *within-platform* audit, the input is systematically manipulated, to study how the resulting outputs differ. For example, in a test for racial bias in Google AdSense, Sweeney (Sweeney 2013) conducted Web searches on names, manipulating them by their racial associations. She then analyzed the content of ads chosen by the algorithm and found that searches on names commonly given to Black children (e.g., DeShawn or Jermaine) were significantly more likely to be served up ads related to arrest, as compared to searches made on white-sounding names (e.g, Emma, Jill).

A second approach, *cross-platform* auditing, can be useful for detecting cases where a system is generally biased (i.e., biases all inputs, rather than only certain categories of input). In their audit of the hotel rating system at Booking.com, (Eslami et al. 2017) compared the ratings of a random set of hotels at Booking, to those at two other platforms. We use both approaches in our study of image taggers. First, using an input set of people images balanced for gender and race,

we examine whether a given algorithm's output tags differ significantly, as a function of the depicted person's gender and race. Secondly, we compare the tags used to describe people across the six taggers, as illustrated in Table 1.

Algorithmic bias and fairness

While the above approach offers a means to systematically examine algorithmic taggers, a more difficult question is how to define and measure the extent to which they exhibit undesirable behavior. "Algorithmic bias" is a term used frequently, yet it is not often defined precisely. Drawing from statistics, a "biased algorithm" would yield outputs that deviate systematically from what we would expect (i.e., the expected value). For instance, in Eslami and colleague's study, the hotel rating algorithm was biased, because the scores were inflated as compared to those input by the users, as well as ratings at other platforms.

In filtering algorithms on the Web or social media, bias is typically understood as a skewing of what is presented to users, as compared to an underlying distribution (i.e., over- or under-representing certain types of media artifacts). Chakraborty and colleagues (Chakraborty et al. 2017) showed that Twitter trends exhibit demographic biases; influencers, the people responsible for making content popular on Twitter, are not representative of the site's overall population of users. Ribeiro and colleagues (Ribeiro et al. 2018) focused on detecting the ideological biases of news sources, in order to ensure that on social media outlets like Facebook or Twitter, the political biases in news is transparent to users. Kay and colleagues (Kay, Matuszek, and Munson 2015) demonstrated the use of offline labor statistics as a baseline, in order to measure gender bias in image search engines, in terms of their representations of the professions.

Researchers aim to detect biases in algorithmic processes, to make them more transparent to users, as well as to promote fairness. However, as noted by Binns (Binns 2018), fairness "is best understood as a placeholder term for a variety of normative egalitarian considerations." In contrast to the above examples, image tagging algorithms do not have an obvious baseline for comparison. What words would we expect a tagger to use, in describing a person? How can we determine if its behavior is socially just?

Computer vision: social biases

There is a growing body of work on social biases in computer vision, which sheds some light on the above questions. With respect to face recognition in images, Klare and colleagues (Klare et al. 2012) found that particular demographic groups (people of color, women, and younger people aged 18 to 30) were more prone to recognition errors than others. Thus, they put forward solutions for biometric systems used in intelligence and law enforcement applications. More recent work attempted to mitigate such biases by applying deep learning (convolutional neural networks - CNNs) in classifying age, gender and ethnicity simultaneously (Das, Dantcheva, and Bremond 2018). However, others claim that while CNNs have brought remarkable improvements in general image recognition tasks, "accurately estimating age and gender in unconstrained settings...remains unsolved." (Levi and Hassner 2015)

In another recent work, a specific gender bias was reported in the popular MS-COCO dataset. Zhao and colleagues (Zhao et al. 2017) found that certain activity labels (e.g., verbs like cooking or shopping) were systematically associated with images of women, while other verbs (e.g., driving) more frequently described images depicting men. To prevent biases in the resulting image tagging models, they developed techniques to constrain the corpus.

Based on the above findings, we examine the following, to shed light on how image tagging APIs interpret people, and whether they do so in a manner that is "fair":

- Gender-related tags, used correctly/incorrectly
- The use of tags judging physical attractiveness
- Tags that make inferences about people's profession or role, emotional state, or character traits.

We consider whether there are systematic differences in the above, as a function of the depicted person's gender or race (within-tagger variation), as well as cross-tagger differences in the use of the three categories of tags.

Methodology

To conduct tagger audits, we generated a set of descriptive tags for all images in the Chicago Face Database, using the six tagging APIs. Considering the entire set of tags, we conducted a thematic analysis, to derive a set of underlying concepts that taggers use when describing people images. Using these concepts, we conducted appropriate statistical analyses to understand within-tagger differences, in terms of how the tool "perceives" men and women of different races, and cross-tagger differences. Our *Social B(eye)as Dataset* (Barlas et al. 2019) can be downloaded from Dataverse[16].

Data collection

We used the image tagging APIs provided by Amazon, Clarifai, Google, Imagga, Microsoft and IBM Watson, using their pre-trained (i.e., general) models. These APIs represent a collection of tools that can be easily used by any developer, without previous knowledge of machine learning. In particular, we executed a REST call to each of the six APIs for each of the images contained in the Chicago Face Database (CFD). The CFD is a free resource[17] consisting of 597 high-resolution, standardized images of diverse individuals, between the ages of 18 and 40 years. The dataset is balanced for gender and race, as detailed in Table 2.

Created by psychologists (Ma, Correll, and Wittenbrink 2015), 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. Using images collected in the wild would challenge our study, as we would

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

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

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

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

surely collect images of varying quality as well as those taken in a large range of social and physical contexts. In other words, by using the CFD images as input to the taggers, we can gauge their treatment of men and women of different races, in a more controlled manner.

Typology of descriptive tags

We aimed to create a typology, which maps the descriptive tags to a set of common concepts. Given that taggers use different vocabularies, the typology helps us understand the manner in which they perceive people, by characterizing which facets of the images are described. To this end, we applied an inductive thematic analysis to the set of tags produced across all six taggers (Herring 2009).

A grounded approach was used, in which three researchers independently inspected the full set of output tags, grouping them by related concepts. The researchers met regularly to discuss the emerging themes, until a consensus was reached. Tags could belong to more than one concept cluster; for instance, the tag "girl" conveys both the gender and relative age of the individual being described. The analysis resulted in four "super-clusters" of tag themes:

- **Demographic:** Tags that describe the inferred gender, age or race of the depicted person.
- **Concrete:** Tags that describe directly observable attributes of the image or the depicted individual.
- **Abstract:** Tags describing attributes of the person that are not directly observable and are based on inference.
- **Other:** Tags that do not have an obvious meaning given the content of the image.

The 16 thematic clusters, and their positioning within the four super-clusters, are depicted in Table 3. Several interesting differences between the taggers can be observed here. Clarifai has the largest vocabulary of tags in our dataset, with tags across all four super-clusters. This can be contrasted to Amazon's Rekognition, which produced the smallest vocabulary of unique tags. As expected, all six taggers use a wide vocabulary of tags to describe concrete observations in the images, such as body parts, or the depicted person's hairstyle and clothing. However, only a subset of taggers, most notably Clarifai, Imagga and Watson, use a rich set of tags that describe abstract, inferred characteristics (e.g., a person's emotional state or physical attractiveness).

To complement the analysis of tagger vocabulary in Table 3, Figure 2 provides a summary of the use of the four super-clusters by the taggers; in particular, the proportion of the images labeled with one or more tags in each category, is detailed. Here we can observe that while Amazon's tagger has a vocabulary for describing demographic characteristics, it used them in labeling only 68% of the CFD images.

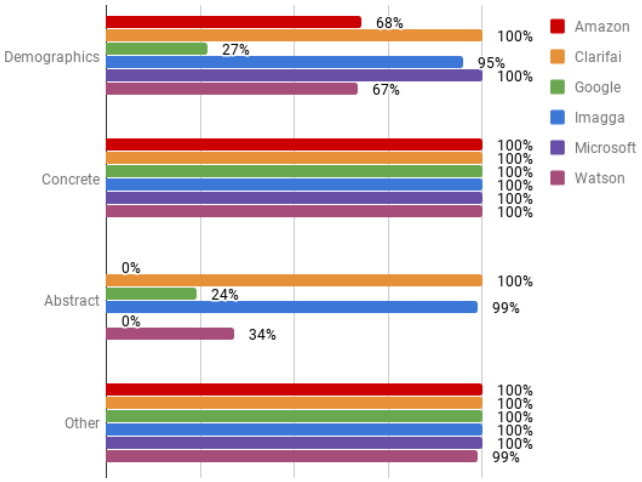

Figure 2: Proportion of images with at least one tag in each super-cluster.

Also, Clarifai, Imagga and Microsoft describe images with a blend of demographic, concrete and abstract tags; Google and Watson use their abstract tags more infrequently.

Table 4 provides a finer-grained comparison between taggers. The output of each tagger on each image was represented as a 16-dimensional vector indicating the counts on each sub-cluster in Table 3. In other words, for a given image description, we counted the number of tags mapping onto each concept. We then normalized these counts by the total number of tags used in the description. For each of the 597 images, we calculated the cosine distance between each pair of taggers description for that image. The mean and median cosine distance are reported in Table 4. Amazon and Google are clearly quite similar in terms of their handling of people images. The other four taggers all have their differences, which we explore further in the analysis.

Analysis

Now we examine how taggers use gender-related tags, based on the person's reported gender and race in the CFD. We then consider taggers' use of judgment tags, which describe a person's perceived attractiveness. Finally, we consider how often taggers describe aspects that are not directly observable - a person's traits, emotional state or professional role.

In our analysis, we consider two measures that quantify the extent to which a tagger uses a given concept when describing people. The first reflects whether or not a concept is used at all to describe a given image (i.e., the proportion of images in the CFD, which contain at least one tag relating to the concept). Given that a concept is used often, we consider a second continuous measure, which reflects the extent to which the concept is prevalent in the tags assigned to a given image (i.e., the mean/median number of tags used that relate to the concept). To examine the correlation of the depicted person's race and gender to the above response variables, we use Logistic Regression (logit) models in the first case, and Analysis of Variance (ANOVA) in the second.

	Example tags	Amazon	Clarifai	Google	Imagga	Microsoft	Watson
Demographics		8	14	6	10	8	18
Gender	boy, girl, masculinity, woman	5	7	4	5	6	11
–Feminine	girl, woman, woman's portrait	4	2	1	1	3	7
–Masculine	boy, man, male child	1	5	3	4	3	5
Age	young, elderly, adult, youth	6	10	5	7	6	15
Race	multicultural	0	1	0	0	0	0
Concrete		23	37	38	31	47	47
Action	staring, looking, smiling, dressed	1	8	3	5	10	0
Body/person	skin, head, face, human	9	7	19	6	7	11
Hair	blond, brunette, mohawk, afro	6	9	13	5	1	9
Clothing	jewelry, dress, bling, makeup	5	3	1	8	11	11
Photo-meta	profile, portrait, studio, indoors	2	6	2	6	7	4
Colors	black, red, green, sage green	1	3	4	2	9	14
Size & Shape	little, long, vertical	0	4	0	0	2	0
Abstract		0	38	2	9	0	14
Judgment	pretty, sexy, attractive, cute	0	6	1	5	0	0
Traits	friendly, serious, crazy, energetic	0	20	0	1	0	0
Emotion	joy, satisfaction, happiness, smiling	0	7	1	1	0	0
Occupation	actor, entertainer, scientist, model	0	7	0	2	0	14
Other	desktop, doughnut, kitchen, temple	1	5	1	4	16	1
Vocabulary size		32	95	48	54	71	75

Table 3: Thematic clusters and respective number of unique tags per API.

	C	G	I	M	W
Amazon	.44/.43	.05/.04	.34/.34	.64/.64	.23/.20
Clarifai		.49/.47	.33/.32	.41/.40	.51/.50
Google			.36/.36	.69/.69	.25/.20
Imagga				.52/.52	.58/.59
Microsoft					.58/.59

Table 4: Mean/median pairwise cosine distance in use of tag themes over all 597 images.

	Neutral	Man	Woman
Amazon	201	67	329
Clarifai	110	359	128
Google	439	148	10
Imagga	131	465	1
Microsoft	150	434	13
Watson	329	141	127

Table 5: Gender inference by taggers (# images).

	Men			Women		
	Prec.	Recall	F_1	Prec.	Recall	F_1
Amazon	1.00	.23	.37	.81	.87	.84
Clarifai	.81	1.00	.89	1.00	.41	.59
Google	.99	.50	.67	1.00	.03	.06
Imagga	.61	.98	.75	1.00	.003	.01
Microsoft	.66	.98	.79	1.00	.04	.08
Watson	.86	.41	.56	.98	.40	.57

Table 6: Precision, recall, F-measure on gender tagging.

Gender (mis)inference

Between-tagger differences. As shown in Table 3, all six taggers have vocabularies that relate to gender. Given the nature of the CFD images, in which one individual is depicted, we considered that a tagger inferred a depicted person to be a man if it used more masculine than feminine tags, in the set of output tags describing a given person, and vice versa. Images that were assigned no gender tags, or equal numbers of masculine/feminine tags, were assumed to be neutral (i.e., no gender inference made by the tagger). Table 5 details the frequency of the inferred genders by taggers, whereas Table 6 presents the precision, recall, and F_1 measure, broken out by the actual gender of the depicted individual.

As can be observed, Google and Watson inferred the gender on fewer than half of the images; therefore, they maintain high precision both when processing images of men and women, but lower recall. In contrast, Imagga labels nearly all images with masculine tags, achieving almost perfect recall (but lower precision) for images of men, and low recall for women. Similarly, Microsoft rarely uses feminine tags. Overall, Clarifai has the most balanced approach, with a relatively high F_1 on both images of men and women.

Within-tagger differences. We now consider whether the accurate use of gender tags is correlated to the depicted persons' race. Table 7 presents estimated Logistic Regression (Logit) models for predicting the event that a tagger has used gender tags correctly, based on the depicted person's reported gender, with his or her race as the explanatory variable. We use the following conventions to report statistical significance: $^{***}\, p < .001$, $^{**}\, p < .01$, $^{*}\, p < .05$. For significant effects, we also report the odds ratio (in parentheses).

For Clarifai and Watson, the likelihood of using gender tags appropriately is lower for Blacks, as compared to Asians, Latinos or Whites. Common facial analysis datasets used for benchmarking algorithms over-represent fair-skinned individuals (Buolamwini and Gebru 2018); thus, we may be observing the results of this bias.

	Intercept	Black	Latino	White
Amazon	.0918	-.0207	.560* (1.75)	.161
Clarifai	1.264***	-0.863** (0.42)	-0.065	-0.341
Google	-1.265***	.291	.774* (2.17)	-.108
Imagga	-.055	-.077	-0.056	-.021
Microsoft	-.055	.025	.055	.109
Watson	-.240	-0.562* (0.57)	0.203	0.053

Table 7: Logit model for predicting correct gender tag use based on race with Asians as reference group.

Judgments on Physical Attractiveness

Between-tagger differences. As detailed in Table 3, three APIs (Clarifai, Imagga, Google) use tags that express judgment about a person's physical appearance. Table 8 details the specific tags used by the tagging APIs, as well as their frequency of use, reported first as the proportion of images on which at least one such tag was used, as well as the mean/median number of judgment tags used per image. Interestingly, it can be seen that Clarifai and Imagga use these types of tags frequently, with Imagga using them to describe almost all images, despite that Clarifai's vocabulary is larger and more diverse. Google has only one tag, "beauty," which it applies in describing only 32 images, all depicting women.

Within-tagger differences. Focusing on Clarifai and Imagga, we now examine whether the extent to which they describe people's physical attractiveness is related to gender and race. Specifically, we consider the proportion of tags output by Clarifai/Imagga for an input image, which are related to judgments on physical attractiveness. We conducted an ANOVA, with gender, race and their interaction as factors. Table 10 reports the relevant F statistics along with significance levels. For significant effects, η^2 (in parentheses) is reported, to gauge the effect size. Finally, for each ANOVA, a Tukey post-hoc test was conducted in order to determine which differences are truly significant.

For both taggers, the depicted person's gender and race are correlated to the use of judgment tags in output descriptions. In particular, images of women tend to be described with more of these tags, as compared to those of men. This finding is perhaps not surprising, given wide attention in society, reflected in both the traditional and new media, to women's physical appearance (Dill and Thill 2007). Furthermore, race is highly correlated to the degree to which a person image will receive attractiveness tags. In Clarifai's output, Blacks were described with fewer such tags than images of others. In Imagga's descriptions, images of Asians received more of these tags than other racial groups.

Emotion tags

Between-tagger differences. Table 11 details the extent to which taggers use words that convey emotions or states, character/personal traits, and one's profession or societal role. As previously emphasized, none of these characteristics are directly observable in the photo; they are inferred by the tagger. With respect to tags reflecting emotion, Clarifai uses these to describe two-thirds of the CFD images, while they are not as frequently used by Google or Imagga.

Within-tagger differences. We explored a binary variable, whether or not an emotion tag is used in a given description, and its relation to the depicted person's gender and race. Table 9 details the respective logit model for each tagger. Clarifai uses more emotion tags when describing images of Asian men as compared to Asian women. Google uses more emotion tags when describing White women as compared to Asian women.

Character or personal traits

Only the Clarifai tagger uses words that infer a depicted person's traits. Table 13 presents an ANOVA to explore the relationship of an individual's race and gender to the use of trait tags. Images of men are described more often with trait tags, as compared to those depicting women. In addition, the effect of race is significant, with Asians being described with the fewest traits.

Occupation or role

Between-tagger differences. Only Watson and Clarifai use tags that infer the depicted person's professional or social role. As shown in Table 11, the use of these tags is relatively rare in Clarifai. In contrast, Watson used at least one such tag in describing a third of the CFD images.

Within-tagger differences. Table 12 details the logit model for predicting the presence of tags related to professional role, in the output tags for a given image. In Clarifai's output, race and gender are not correlated to the use of occupation tags. In contrast, Watson's behavior reflects a significant effect on race, with Black men receiving few tags as compared to Asian women.

Discussion

Image analysis algorithms have transformed the way that we interact with information and other people, from facilitating the organization and sharing of large multimedia collections, to enhancing personalization and interactivity in applications across domains. Over the past years, it has become increasingly easy for any developer - regardless of her experience with or knowledge of machine learning - to have access to state-of-the-art AI modules as Cognitive Services.

It must be noted that just as the new "Algorithmic Economy" is fueling innovation, it is also having an impact on socio-technical research. Even before the rise of image analysis APIs, researchers had begun implementing published computer vision algorithms to conduct large-scale analyses of visual communication on social media (e.g., Hu and colleagues' study of the content of images shared on Instagram (Hu et al. 2014)). However, commercial, pay-for-use image analysis APIs have become extremely popular in recent years with computational social scientists and those working in human-computer interaction. For instance, (Garimella,

	Tags	Prop. Images	Mean/Median
Clarifai	attractive, cute, fine looking, pretty, sexy, strange	0.72	0.05/0.05
Google	beauty	0.05	0.005/0
Imagga	attractive, cute,handsome, pretty, sexy	0.98	0.13/0.13

Table 8: Judgment tags and frequency of use.

	Intercept	Men	Black	Latino	White	MB	ML	MW
Clarifai	0.941**	1.300* (3.67)	-0.219	-0.506	0.0146	-0.291	0.506	-0.896
Google	-1.966***	-0.275	0.868	0.440	1.222** (3.40)	0.141	-1.418	-0.799
Imagga	-0.856**	-18.710	0.133	0.025	-0.676	-0.134	-0.025	15.720

Table 9: Logit model for predicting use of emotion tags with Asian Women as reference group.

	Gender	Race	G*R	Sig. diff
Clarifai	531.8*** (0.44)	28.0*** (0.07)	4.2** (.01)	G: W>M R: A,L,W>B
Imagga	94.5*** (0.13)	9.5*** (0.04)	5.6** (0.02)	G: W>M R: A>B,L,W

Table 10: ANOVA: use of judgment tags.

Alfayad, and Weber 2016) used Imagga to analyze the content of images on Instagram to glean information about public health, in the spirit of Google Flu Trends.[18]

Going a step further, many researchers use automatic facial analysis. Liu and colleagues (Liu et al. 2016) used two APIs based on deep learning methods, Face++[19] and EmoVu,[20] to infer demographic characteristics and emotion. The goal was to predict a user's personality type based on her Twitter profile picture. Similarly, (Deeb-Swihart et al. 2017) used Face++ in their study, which aimed to identify types of selfies on Instagram, and who tends to post them. Finally, as an example of image analysis APIs being used in a sensitive social context, (Kocabey et al. 2018) used Face++ to infer people's body mass index from profile pictures, to study the relationship between popularity and weight.

This surge in research and development brought about by the successes of machine learning, as in the case of computer vision, has co-occurred with increasing concerns about algorithmic biases and the potential for opaque processes to result in social harm and discrimination. In other words, there is growing awareness that algorithmic processes should be **fair**. As stated by Ekstrand and colleagues in their work on ensuring users' privacy (Ekstrand, Joshaghani, and Mehrpouyan 2018), fairness in socio-technical systems is challenging; the social consequences of systems interact in complex ways with their ethical and legal effects.

While machine learning researchers looking at classification tasks have put forward ways to operationalize fairness, when working with machine output that is more open-ended, it is less clear how to approach the issue. Still, some common operationalizations of fairness might be applied to our results. In particular, the "group fairness" interpretation holds that different groups of people should experience comparable error rates (Feldman et al. 2015). In our analysis, we observed that image analysis taggers' interpretations of people images do differ systematically across gender and racial groups. With respect to the use of gender-related tags, all taggers use gender-related tags when interpreting images of men versus women (Table 5). However, it is telling that the taggers with the most balanced approach, Clarifai and Watson, tended to use gender tags incorrectly when describing images of Black people (Table 7).

Similarly, with respect to the use of tags related to one's physical attractiveness, there are stark differences between social groups. Specifically, the taggers use these words more frequently when describing images of women versus men, reinforcing the social expectations on women to be attractive (Table 10). Certain attractiveness tags that Clarifai uses are also highly gendered, e.g., "cute," "pretty," "sexy" and "attractive" are used exclusively when describing women, while "fine looking" is used to describe both men and women. Even more disturbing is that Clarifai labels images of Black individuals with "attractiveness tags" significantly less often than other social groups.

Finally, with respect to the abstract attributes (traits, emotions and occupations) there were mixed results. However, with respect to Clarifai, it is interesting that while we saw that it was more likely to describe women's physical attractiveness, it was also more likely to comment on men's character traits. For instance, it is quite telling that the tag "intelligence" is applied to 55 images, 53 of which depict men.

In the above examples, "group fairness" is arguably not achieved by the taggers. In future work, we can consider how to score image tagging APIs across a number of defined characteristics. This would enable possible consumers (i.e., developers and researchers) to be aware of the specific scenarios in which the algorithms are less than fair, when processing people-related images. We should also consider users' perceptions of fairness in this open-ended task, and how they might differ depending on the context (e.g., auto-tagging in the context of organizing one's personal photo collection as compared to facilitating search and retrieval of profile images on a dating site).

[18]https://www.google.org/flutrends/about/

[19]http://www.faceplusplus.com

[20]http://emovu.com

	Tags	Prop. Images	Mean/Median
Emotion			
Clarifai	enjoyment, happiness, joy, satisfaction	0.76	0.04/0.05
Google	emotion	0.19	0.02/0
Imagga	happy	0.14	0.02/0
Traits			
Clarifai	attitude, casual, cool, confidence, contemporary, crazy, elegant, energetic, fashionable, friendly, fun, individuality, innocence, intelligence, masculinity, pensive, serious, strange, strength, trendy	0.99	0.18/0.19
Occupation			
Clarifai	athlete, business, military, model	0.07	0.003/0
Watson	actor, anchorperson, careerist, celebrity, entertainer, movie actor, official, operator, orator, representative, scientist, social scientist, thinker, woman orator	0.34	0.05/0

Table 11: Tags describing emotions/traits/occupations and frequency of use.

	Intercept	Men	Black	Latino	White	MB	ML	MW
Clarifai	-2.342***	-0.877	0.561	0.019	-0.726	0.841	-16.36	-15.62
Watson	-0.856***	0.0447	0.740* (2.09)	0.568	0.061	-3.330*** (0.04)	-0.0672	0.643

Table 12: Logit model for predicting use of occupation tags with Asian Women as reference group.

	Gender	Race	G*R	Sig. diff
Clarifai	283.6*** (0.30)	19.4*** (0.06)	3.17* (0.01)	G: M>W R: B,L,W>A

Table 13: ANOVA: use of tags describing traits.

Limitations

We have only considered image tagging APIs and not facial recognition. As mentioned, these represent general tools that developers would be likely to use to process images in social applications, to infer the content of images across domains. Finally, we highlight the fact that the the CFD images are highly controlled, as compared to images shared in the wild. In addition, CFD uses a discrete classification for characterizing the race and gender of depicted persons. There are no persons labeled as mixed race, or as being of non-binary gender; such individuals' images might challenge the taggers, particularly in terms of their use of demographic tags.

Conclusions

Researchers and practitioners should maintain a healthy skepticism of proprietary solutions for the automatic analysis of people-related media. On a positive note, researchers are working toward detecting the roots of social biases in image analysis, such as training data sets that under-represent minorities (Buolamwini and Gebru 2018). In addition, diversity in people image processing is receiving attention, such as in the InclusiveFaceNet initiative (Ryu, Adam, and Mitchell 2018). Thus, future tools may be positioned to treat images of people, across genders and races, more fairly.

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

Barlas, P.; Kyriakou, K.; Kleanthous, S.; and Otterbacher, J. 2019. Social B(eye)as: Human and Machine Descriptions of People Images. In *Proceedings of the 13th Annual Conference on Web and Social Media*, ICWSM '19. AAAI.

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

Birnholtz, J.; Fitzpatrick, C.; Handel, M.; and Brubaker, J. R. 2014. Identity, identification and identifiability: The language of self-presentation on a location-based mobile dating app. In *Proceedings of the 16th international conference on Human-computer interaction with mobile devices & services*, 3–12. ACM.

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

Burrell, J. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society* 3(1):2053951715622512.

Chakraborty, A.; Messias, J.; Benevenuto, F.; Ghosh, S.; Ganguly, N.; and Gummadi, K. P. 2017. Who makes trends? understanding demographic biases in crowdsourced recommendations. In *ICWSM*, 22–31.

Chua, T. H. H., and Chang, L. 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* 55:190–197.

Das, A.; Dantcheva, A.; and Bremond, F. 2018. Mitigating bias in gender, age and ethnicity classification: a multi-

task convolution neural network approach. In *ECCVW 2018-European Conference of Computer Vision Workshops*.

Deeb-Swihart, J.; Polack, C.; Gilbert, E.; and Essa, I. A. 2017. Selfie-presentation in everyday life: A large-scale characterization of selfie contexts on instagram. In *ICWSM*, 42–51.

Diakopoulos, N. 2016. Accountability in algorithmic decision making. *Communications of the ACM* 59(2):56–62.

Dill, K. E., and Thill, K. P. 2007. Video game characters and the socialization of gender roles: Young people's perceptions mirror sexist media depictions. *Sex roles* 57(11-12):851–864.

Ekstrand, M. D.; Joshaghani, R.; and Mehrpouyan, H. 2018. Privacy for all: Ensuring fair and equitable privacy protections. In *Conference on Fairness, Accountability and Transparency*, 35–47.

Eslami, M.; Rickman, A.; Vaccaro, K.; Aleyasen, A.; Vuong, A.; Karahalios, K.; Hamilton, K.; and Sandvig, C. 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*, 153–162. ACM.

Eslami, M.; Vaccaro, K.; Karahalios, K.; and Hamilton, K. 2017. " be careful; things can be worse than they appear": Understanding biased algorithms and users' behavior around them in rating platforms. In *ICWSM*, 62–71.

Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259–268. ACM.

Garimella, V. R. K.; Alfayad, A.; and Weber, I. 2016. Social media image analysis for public health. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 5543–5547. ACM.

Gillespie, T. 2014. The relevance of algorithms. *Media technologies: Essays on communication, materiality, and society* 167.

Herring, S. C. 2009. Web content analysis: Expanding the paradigm. In *International handbook of Internet research*. Springer. 233–249.

Hu, Y.; Manikonda, L.; Kambhampati, S.; et al. 2014. What we instagram: A first analysis of instagram photo content and user types. In *ICWSM*.

Kay, M.; Matuszek, C.; and Munson, S. A. 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*, 3819–3828. ACM.

Klare, B. F.; Burge, M. J.; Klontz, J. C.; Bruegge, R. W. V.; and Jain, A. K. 2012. Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security* 7(6):1789–1801.

Kocabey, E.; Ofli, F.; Marin, J.; Torralba, A.; and Weber, I. 2018. Using computer vision to study the effects of bmi on online popularity and weight-based homophily. In *International Conference on Social Informatics*, 129–138. Springer.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

Levi, G., and Hassner, T. 2015. Age and gender classification using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 34–42.

Liu, L.; Preotiuc-Pietro, D.; Samani, Z. R.; Moghaddam, M. E.; and Ungar, L. H. 2016. Analyzing personality through social media profile picture choice. In *ICWSM*, 211–220.

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

O'Neil, C. 2017. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

Ribeiro, F. N.; Henrique, L.; Benevenuto, F.; Chakraborty, A.; Kulshrestha, J.; Babaei, M.; and Gummadi, K. P. 2018. Media bias monitor: Quantifying biases of social media news outlets at large-scale. In *ICWSM*, 290–299.

Ryu, H. J.; Adam, H.; and Mitchell, M. 2018. Inclusivefacenet: Improving face attribute detection with race and gender diversity. In *Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)*.

Sandvig, C.; Hamilton, K.; Karahalios, K.; and Langbort, C. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry* 1–23.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Stefanone, M. A.; Lackaff, D.; and Rosen, D. 2011. Contingencies of self-worth and social-networking-site behavior. *Cyberpsychology, Behavior, and Social Networking* 14(1-2):41–49.

Sweeney, L. 2013. Discrimination in online ad delivery. *Queue* 11(3):10.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–9.

Wilson, M. 2017. Algorithms (and the) everyday. *Information, Communication & Society* 20(1):137–150.

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