Shifting Our Awareness, Taking Back Tags: Temporal Changes in Computer Vision Services' Social Behaviors

Pınar Barlas,[1] Maximilian Krahn,[1,2] Styliani Kleanthous,[1,3]
Kyriakos Kyriakou,[1,3] Jahna Otterbacher[1,3]
[1]CYENS Centre of Excellence (Cyprus)
[2]Aalto University (Finland)
[3]Cyprus Center for Algorithmic Transparency, Open University of Cyprus (Cyprus)
p.barlas@cyens.org.cy; maximilian.krahn@aalto.fi; {s.kleanthous; k.kyriakou; j.otterbacher}@cyens.org.cy

Abstract

Much attention has been on the behaviors of computer vision services when describing images of people. Audits have revealed rampant biases that could lead to harm when services are used by developers and researchers. We focus on temporal auditing, replicating experiments originally conducted three years ago. We document the changes observed over time, relating this to the growing awareness of structural oppression and the need to align technology with social values. While we document some positive changes in the services' behaviors, such as increased accuracy in the use of gender-related tags overall, we also replicate findings concerning larger error rates for images of Black individuals. In addition, we find cases of increased use of inferential tags (e.g., emotions), which are often sensitive. The analysis underscores the difficulty in following changes in services' behaviors over time, and the need for more oversight of such services.

Introduction

The broad area of Fairness, Accountability, Transparency and Ethics (FATE) in data-driven AI has flourished in recent years. In addition to the emergence of dedicated research communities,[1] FATE has become a key topic of interest within established communities such as ICWSM, where researchers aim to uncover how algorithmic bias might harm users of social platforms (e.g., Asplund et al. 2020; Ye, You, and Robert Jr 2017), develop more fair techniques for analyzing social media (e.g., Zeng et al. 2021) or examine the ethical implications of using social media analysis to inform decisions affecting the public (e.g., Mashhadi et al. 2021).

Documenting the social biases exhibited by the *algorithmic tools* used by researchers, and which make up the infrastructures used to study large-scale communication behaviors, is an issue of particular relevance to ICWSM (Jung et al. 2018). In the current work, we focus on image tagging algorithms (ITAs) which infer what is depicted in an input image in the form of output text labels (hereon: tags). These ITAs are offered as paid Cognitive Services for developers to integrate into their applications, providing new functionalities to their end-users. The Software-as-a-Service (SaaS) approach represents a growing practice, as it helps developers and researchers enhance their productivity in an economical way. Although these services are widely used in our information ecosystem, many have expressed concerns about their behavior that is often found to be unfair and/or biased against certain social groups.

This concern is highly reflected in the scientific literature. In their work *"Gender Shades,"* Buolamwini and Gebru found accuracy disparities in commercial gender classification, stating that all classifiers performed best for lighter individuals and males overall, but they also performed worst for darker females (Buolamwini and Gebru 2018). Likewise, while the performance of facial recognition algorithms has improved over time, they still tend to perform best on images depicting white men.[2] Inspired by Gender Shades, others have investigated the performance disparities for race, gender and age in computer vision during the past years, highlighting higher error rates for specific race/gender identities (Kyriakou et al. 2019), skin colors (Raji and Buolamwini 2019), and age groups (Phillips et al. 2011).

The importance of auditing as a means to provide oversight of algorithms has gained attention over the past few years (Taddeo and Floridi 2018; Rahwan 2018; Metaxa et al. 2021). But beyond the efforts of the scientific community, civil society is also questioning the behavior of algorithms, as our understanding of social bias and structural oppression evolves. "Everyday" incidents reflecting social bias are frequently discussed in the news or via social media, particularly those involving products and services from the tech giants, which clearly highlight the need for oversight.

One of the most problematic examples of social bias in computer vision is 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."[3] In a similar incident in September 2021, Facebook issued an apology when its recommendation engine mistakenly suggested to a number of

[1]See, for instance, ACM/AAAI AIES, ACM FAccT, or CAIP, to name just a few.

[2]https://www.nature.com/articles/d41586-020-03186-4

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

users, who watched a newspaper video featuring Black men, if they wanted to *"keep seeing videos about primates"*.[4]

After much work on gender-related algorithmic bias from various researchers, Google announced in early 2020 the removal of all *gender labels* from their Cloud Vision API[5], stating that *it was not possible to infer someone's gender solely from their appearance.*[6] A few months later, in June 2020, IBM decided to discontinue its facial recognition service for *"mass surveillance or racial profiling"*.[7] As will be detailed, its ITA service was later discontinued as well.

Given these observed changes, we focus on *temporal auditing*, aiming to characterize the particular changes in ITAs over the past three years. Beyond those disclosed by big tech, we investigate in parallel whether other services made important (but perhaps subtle) updates they didn't announce, such as dropping the use of particular descriptive tags, in order to enhance accuracy or make their services more socially sensitive. As will be explained, we consider the global changes in the ITAs' behaviors across two points in time, nearly three years apart, as well as changes in the usage of particularly sensitive tags.

Background

Algorithmic auditing started as a method for revealing biases that were systematically emerging in widely used software applications (Bandy 2021). In an attempt to understand what was/is the real impact of these biases in society, researchers acted as third party auditors, exposing how major software platforms were systematically discriminating against certain social groups (Raji and Buolamwini 2019). One of the first algorithm audits of a big tech service found that the representation of women in Google image search results systematically differed as compared to metrics from the U.S. Bureau of Labor Statistics (Kay, Matuszek, and Munson 2015). Race and gender biases were among the most frequent cases of bias detected that can have a real impact at the societal and personal level (Otterbacher, Bates, and Clough 2017; Chen, Johansson, and Sontag 2018). For example, an audit found that some emotion analysis services (EAS) using computer vision were perpetuating emotion stereotypes based on race; the EAS were more likely to infer anger in photos of Black individuals in different cases, which is similar to the psychological tendency to categorize ambiguous faces of Black individuals under emotions of hostility (e.g., anger) (Kyriakou et al. 2020).

A recent audit on computer vision (tagging) platforms consisted of a controlled dataset of people images, imposed on gender-stereotyped backgrounds (Barlas et al. 2021a). Evaluating five proprietary algorithms, the authors found that three of those were misgendering the depicted person when a background was introduced. In an audit of facial recognition platforms (Klare et al. 2012), results showed algorithms to have consistently lower accuracy for women, Black individuals, and younger ages (18-30) compared to the remaining groups within their demographic. A different issue was revealed when Western versus Eastern origin algorithms were tested in face recognition (Phillips et al. 2011), with the Western algorithms recognizing Caucasian faces more accurately than East Asian faces and vice versa.

Algorithmic auditing can help shed light into the issues with computer vision; however, dataset audits are equally important in this context, since computer vision algorithms are trained on image data collected and/or annotated by humans. Analyzing the person subtree in ImageNet, (Yang et al. 2020) revealed gender, race, age and ethnicity inequalities. This is of great concern given that ImageNet is a popular source of training data for many vision algorithms. Acknowledging the underrepresented social groups in image datasets, Karkkainen and Joo have developed a "more balanced" dataset that produces better accuracy across races, compared to other datasets (Karkkainen and Joo 2021).

Finding the root cause of the problem in an opaque algorithmic system is extremely complex, perhaps even impossible, as we have observed in the examples presented in the Introduction. However, auditing to reveal social biases and misbehaviors of a system can act as leverage to encourage more social responsibility by their owners/developers, as well as for raising users' awareness of bias and stereotyping in algorithmic systems they interact with daily. In short, there is a need for more rigorous approaches in auditing algorithmic systems for understanding their behavior, identifying potentially harmful, discriminatory, unfair and/or biased output (Bandy 2021). Referring back to Google's announcement to remove gender-related tags from Cloud Vision, the company explained that this is an example of "how technology should evolve alongside cultural understanding".[8] Thus, *temporal auditing* is an essential tool for gauging the extent to which an algorithmic system or service is evolving, alongside societal changes and social awareness.

Motivation and Research Questions

The importance of replication in social computing research has long been discussed (e.g., Hornbæk et al. 2014), and algorithm audits are no exception. In their work (Raji and Buolamwini 2019) replicate the Gender Shades study, revealing improvements in many of the problematic cases that the first study disclosed, but point at still unresolved issues. In a study of temporal sensitivity in crowdwork, (Christoforou, Barlas, and Otterbacher 2021) replicated an image annotation task at two points in time, 18 months apart, with the first point being before the COVID-19 pandemic and the second point during the height of the pandemic, alongside the social unrest surrounding racial discrimination in the U.S. When describing people images during the pandemic and social unrest, the U.S.-based workers were more likely to describe aspects of the depicted person's identity, as well as their body weight, as compared to 18 months previously. (Metaxa et al. 2021) in a series of studies revisited the ex-

[4]https://www.bbc.com/news/technology-58462511

[5]https://diversity.google/story/ethics-in-action-removing-gender-labels-from-clouds-vision-api/

[6]https://www.businessinsider.com/google-cloud-vision-api-wont-tag-images-by-gender-2020-2

[7]https://www.bbc.com/news/technology-52978191

[8]https://diversity.google/story/ethics-in-action-removing-gender-labels-from-clouds-vision-api/

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

Figure 1: LF-200 (left) and BM-026 (right) from the CFD.

periment run by Kay, Matuszek, and Munson to compare the representation of women in Google's image search results in 2020 to 2015, finding little improvement during that period.

Following this line of thought, as well as the shifting social awareness, we are motivated to ask: *Given the increasing awareness of the social biases in Computer Vision over the past three years, how do the behaviors of the six ITAs compare to those we observed back in 2018?* Thus, we pose the following research questions (RQs):

- RQ1: How did the ITA vocabulary change over time?
- RQ2: How did the use of the tags by the ITAs, with respect to the social groups, change over time?

After answering the RQs, we discuss whether the temporal changes of the ITAs reflect the societal changes we have observed in the years between the two audits.

Methodology

We replicated the data collection (Barlas et al. 2019) and analysis (Kyriakou et al. 2019) followed in October 2018, this time in August 2021. We compared the two analyses, aiming to uncover the differences in the behaviors of the ITAs after almost three years.

Input Images

We used the 597 standardized images from the Chicago Face Database[9] (CFD) (Ma, Correll, and Wittenbrink 2015) depicting individuals between the ages of 18 and 40 years, balanced for their self-reported gender and race as detailed in Table 1. The CFD was chosen for our initial experiment in 2018 as it depicts diverse individuals in a similar, controlled manner with neutral facial expressions, maximizing the probability that the outputs will reflect differences in how different social groups are treated by the ITAs. Examples of the images can be seen in Figure 1.

Image Tagging Algorithms

Using the 597 images from the original CFD, we queried the same six Image Tagging Algorithms (ITAs) that we did in 2018. However, we found that one service – IBM's Watson – was set to be discontinued at the end of December 2021, and that as of 7 January 2021, the service does not allow creating new instances.[10,11] Therefore, our analyses in 2021 include the five services which are still offering their ITAs: **Amazon** Rekognition Image,[12] **Clarifai**,[13] **Google** Cloud Vision,[14] **Imagga** Auto-tagging,[15] and **Microsoft** Computer Vision.[16]

As in 2018, with the ITAs offering specific models (e.g., food or celebrity recognition), we opt for the "General" models that claim to "recognize [...] different concepts including objects, themes, moods, and more", providing "a great all-purpose solution for most visual recognition needs".[17]

The outputs were processed in the same manner as our 2018 pipeline: we tokenized the tags, replaced the space (" ") in multi-word tags with underscore ("_"). New tags appearing in 2021 were analyzed and placed into our thematic typology, thus updating our dictionaries of tags that correspond to each theme. We present the new tags, along with a brief description of the typology, in the following section.

Analysis & Findings

How Did the ITA Vocabulary Change Over Time?

To answer RQ1, we consider both quantitative and qualitative changes in the vocabulary of tags of the five ITAs. We examine the new vocabulary that emerged in 2021, which did not appear when describing the CFD images in 2018, and whether these new tags fit into our previous typology. Likewise, we investigate whether there are any tags that appeared in 2018, but do not appear in 2021 despite descibing the same dataset. Although we discuss seeing "new" tags in 2021 that we didn't in 2018, and failing to find some of the tags in 2021 that we previously found in 2018, our results do not necessarily mean that the services "added" or "discarded" these tags to/from their overall vocabulary. It is possible that the model was updated in such a way that it no longer "sees" the tag in any of the CFD images, but still uses the tag for other images. Similarly, it is possible that "new" tags are in fact tags that were part of the vocabulary of the service in 2018, but did not appear for our study with the CFD. We use terms such as "added", "new tags" and "discarded" to indicate that these tags were not observed in our results at one point.

Typology of Tags. For our 2018 audit, we manually constructed a typology of tags which allowed us to compare the taggers in their use of concepts, despite their differing vocabularies. While our paper (Kyriakou et al. 2019) and the

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

[10]https://cloud.ibm.com/apidocs/visual-recognition

[11]https://cloud.ibm.com/docs/visual-recognition?topic=visual-recognition-release-notes&locale=en#1december2020

[12]https://aws.amazon.com/rekognition/image-features/

[13]https://www.clarifai.com/models/image-recognition-ai

[14]https://cloud.google.com/vision?hl=en#section-3

[15]https://imagga.com/solutions/auto-tagging

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

[17]https://www.clarifai.com/models/image-recognition-ai

Cluster	Example Tags	Amazon		Clarifai		Google		Imagga		Microsoft	
Demographics		8	10	14	13	7	0	10	11	8	8
Feminine	girl, woman, lady	4	4	2	2	1	0	1	2	3	3
Masculine	boy, man, guy	1	2	5	5	3	0	4	4	3	3
Age	young, elderly, child	6	8	10	9	6	0	7	8	6	6
Race	multicultural	0	0	1	1	0	0	0	0	0	0
Concrete		23	34	38	33	36	60	31	28	47	41
Action	staring, wear, laughing	1	2	7	7	2	5	5	4	10	10
Body/Person	eyes, face, human	9	12	7	9	19	23	6	7	7	6
Hair	hair, blonde, afro	6	4	9	9	12	14	5	5	1	1
Clothing	apparel, sweater, lipstick	5	12	3	1	1	11	8	7	11	11
Photo-meta	portrait, indoors, mugshot	2	4	8	4	2	4	6	5	7	7
Colors	black, green, dark	1	2	3	4	4	5	2	1	9	4
Size & Shape	large, long, curly	0	1	4	2	1	1	0	0	2	2
Abstract		0	2	38	32	3	2	10	12	0	0
Judgment	pretty, sexy, cute	0	0	7	6	1	0	5	6	0	0
Traits	friendly, serious, casual	0	0	20	17	0	0	1	1	0	0
Emotion	joy, angry, smile	0	1	8	9	2	2	2	2	0	0
Occupation	performer, model, son	0	1	5	3	0	0	2	3	0	0
Other	desktop, doughnut, temple	1	1	9	3	2	7	4	8	16	13
Vocabulary size		32	46	95	77	47	66	54	57	71	62

Table 2: Number of tags falling into each cluster in 2018 (left) and 2021 (right). Cells showing a net change are highlighted.

	2018 only	**2021 only**	**Net change**
Amazon	10	24	+14
Clarifai	30	12	-18
Google	17	35	+18
Imagga	7	10	+3
Microsoft	9	0	-9

Table 3: # of tags that appeared in 2018 but not in 2021 ("2018 only"), # of tags that appeared in 2021 but not in 2018 ("2021 only"), and the net change in vocabulary.

accompanying dataset (Barlas et al. 2019) have more information, we briefly describe the typology here. The typology consists of four super-clusters, which in turn contain a total of 15 sub-clusters. The clusters and their relations can be seen in Table 2, along with example tags from each subcluster. It is important to note that the sub-cluster names are shortened and simplified for convenience, but in fact house additional, related concepts.

Table 2 shows the number of tags that fall into each sub- and super-cluster in the two datasets of output tags (2018 v. 2021) we are investigating. It is important to note a few things, while interpreting these numbers. One is that the clusters are not mutually-exclusive; one tag might contain multiple meanings, and as such can fall into more than one cluster. An example is "girl", which indicates both a Feminine gender and an Age. Another fact to keep in mind is that these numbers are the *net change* in tags; while an ITA may seem like it "added" two tags to a cluster (seeing a change of +2 from 2018 to 2021), it is possible that the total number of new and discarded tags are far greater than this number. For example, Clarifai appears to have only one additional Emotion tag in 2021, but a closer look (e.g., Table 4) shows that there were in fact three additions to and two removals from the vocabulary, resulting in a *net increase* of 1.

Overall, two ITAs substantially increased their vocabulary size (by approximately 40%), one ITA remained approximately the same (only 6% growth), while two decreased (19% & 13%). Amazon changed its vocabulary in almost every subcluster, with all but one change in the positive direction. As such, Amazon has the biggest growth in vocabulary relative to its total size (+44%), despite still having the smallest vocabulary. Google, while removing all Demographic and Judgment tags, still added more tags than Amazon; however with a bigger overall vocabulary, the relative growth is smaller (+40%).

Clarifai, while having a net decrease in the total vocabulary, still has the biggest vocabulary at 77 unique tags. Lastly, Imagga was the ITA with the least changes in total vocabulary (total unique tags in 2021 is only 3 more than in 2018), although Microsoft changed fewer absolute number of tags (9 tags) (see Table 3).

Demographics. Demographic tags can be sensitive, depending on the context in which the ITA will be used and, of course, the accuracy of the ITA's gender inferences (e.g. Buolamwini and Gebru 2018; Scheuerman, Paul, and Brubaker 2019). Google stopped using gender tags, as reported. In addition, the age-related tags were also not observed – no demographic tags were used by Google for the images in our 2021 audit.

Clarifai only discarded one Age tag ("youth"); other Demographic tags remained unchanged. Interestingly, Clarifai remains the only ITA to use a race/ethnicity-related tag ("multicultural"). Amazon on the other hand, is the only ITA that *added* Demographic tags. The tags newly observed were "boy" (Masculine & Age) and "teen" (Age). Imagga and Mi-

Subcluster	Tag changes
Judgment	attractive (-); strange (-); beautiful (+)
Emotion	enjoyment (-); satisfaction (-); angry (+); relaxation (+); cheerful (+)
Traits	confidence (-); fashionable (-); individuality (-); innocence (-); strength (-); charming (+); cheerful (+)
Occupation	athlete (-); business (-); military (-); scholar (+)

Table 4: Clarifai's changes in the Abstract cluster. (+) indicates a tag was seen in 2021 but not 2018; (-) vice versa.

crosoft had no change in their Demographics tags at all.

Abstract. The Abstract tags represent attributes for which there is no visual evidence in the input photo. In other words, these tags are inferential in nature and could therefore also be sensitive. Microsoft, as in 2018, used no Abstract tags. Amazon, which previously had used no Abstract tags, in 2021 used one new Emotion tag ("smile") and one new Occupation tag ("performer"). Imagga slightly grew its Abstract vocabulary as well, adding one Judgment ("smasher") and one Occupation tag ("cover girl").

Google added one Emotion tag ("happy") and removed another Emotion tag ("emotion"), keeping the subcluster total stable. In addition, the only Judgment tag Google previously had ("beauty") was not observed in the 2021 outputs.

Clarifai, which previously boasted 38 Abstract tags, reduced its total number of unique tags in the supercluster. However, the net change comes from multiple tags being removed while some others were added; the specific vocabulary changes can be seen in Table 4.

Concrete & Other. Amazon and Google grew their Concrete vocabulary significantly. Both introduced a lot of new Clothing tags (although some were discarded, such as Amazon's "bling" tag). Google added some Hair tags, e.g. "cornrows". Amazon on the other hand primarily discarded Hair tags, including "afro_hairstyle" and "mohawk".

Body/Person was another subcluster with interesting changes. Amazon and Google introduced some new tags, the latter adding tags such as "human_body", "mammal", and "vertabrate". Clarifai did not use the "no_person" tag (which it had used twice in our 2018 audit), while Google took away the "person" tag.

Vectorization. Given that some (or most) of the tags used by each ITA per image are the same in 2018 and in 2021, we decided to use the TF-IDF embedding method (Salton and Buckley 1988) to convert our tags into vectors in euclidean space, where each unique tag was assigned one dimension. The set of tags output for each image by each ITA in 2018 was thus converted into a single vector, which was then compared to its equivalent in the 2021 audit, such that the cosine distance of these vectors show the (dis)similarity of the ITA's tagging behavior per image, in 2018 and 2021. Table 5 presents the mean/median distance of each ITA's results from 2018 and 2021. Google and Microsoft appear to have the least amount of change, remaining very similar to their tagging behavior of 2018. On the other hand, Amazon and Clarifai changed a fair bit. These results reflect the trend in overall vocabulary changes, which may partially explain the rate of change.

ITA	Distance
Amazon	0.061 / 0.053
Clarifai	0.066 / 0.054
Google	0.014 / 0.008
Imagga	0.047 / 0.036
Microsoft	0.014 / 0.012

Table 5: Mean/Median cosine distance of TF-IDF vectors of each ITA's 2018 and 2021 outputs, across the 597 images.

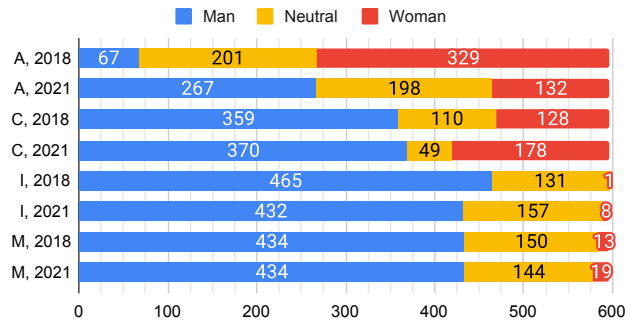

Figure 2: Gender inference by taggers (# images) over all 597 images, in 2018 and 2021.

How Did the Use of the Tags by the ITAs, With Respect to Social Groups, Change Over Time?

Following the analyses performed in 2018, we look at use of the gendered and other abstract tags used by the ITAs.

Gender Inferences. The gendered tags are separated into masculine and feminine, following the binary gender indicated in the CFD metadata. We assume that an ITA interprets a given image as depicting a woman if the proportion of feminine tags is greater than the proportion of masculine tags used to describe it, and vice versa. In the event of a tie or no gendered tags, we assume the ITA's interpretation is neutral. We compare the ITA's interpreted gender to the CFD ground truth, to calculate the precision, recall, and F_1 measure for each ITA's gender inference. The performance is compared across social groups to see if the gender and/or race of the depicted person is correlated to the accuracy of the gender inferences; all results are then compared to those of 2018. Given that Google has removed all gender-related tags from the service, we exclude Google from this analysis.

Figure 2 shows the number of images where each ITA inferred a gender, in 2018 and in 2021. Imagga and Microsoft performed almost exactly as they had in 2018, opting for inferring masculine gender on most images, and feminine gender on very few images. Clarifai generally inferred a gender on more images in 2021, especially increasing the number

	Men			Women		
	Prec.	Recall	F_1	Prec.	Recall	F_1
A 2018	1.00	.23	.37	.81	.87	.84
A 2021	.90	.82	.86	1.00	.43	.60
C 2018	.81	1.00	.89	1.00	.41	.59
C 2021	.78	.99	.87	.99	.58	.73
I 2018	.61	.98	.75	1.00	.003	.01
I 2021	.66	.99	.87	1.00	.03	.05
M 2018	.66	.98	.79	1.00	.04	.08
M 2021	.66	.99	.79	1.00	.06	.12

Table 6: Precision, recall, F-measure on gender tagging for Men (left) and Women (right) per ITA, in 2018 and in 2021.

	Intercept	Black	Latino	White
Amazon	$.6789^{***}$	-0.2118^{***} (0.809)	.0433	.0150
Clarifai	0.8440^{***}	-0.1283^{**} (0.878)	-0.0107	-0.0735
Imagga	$.5229^{***}$	-.0559	-0.0322	-.0202
Microsoft	.532	-.030	-.0228	-.0239

Table 7: Logit model for predicting correct gender tag use based on race with Asians as reference group, in 2021.

of images for which it inferred a feminine gender. Amazon, on the other hand, greatly reduced its feminine gender inferences as compared to 2018, and instead inferred a masculine gender much more often in 2021.

Table 6 compares the accuracy on gender inferences in the 2018 and 2021 audits. As expected from Figure 2, Amazon's accuracy in using gender tags appropriately when describing images of men increased over time, whereas its accuracy on images of women has decreased. Clarifai appears to have improved its performance in 2021, maintaining nearly the same F_1 on images of men, and improving its F_1 when analyzing images of women. Similarly, Imagga improved over time in describing images of men with correct gender tags, while remaining quite poor in describing women. No substantial changes are observed for the Microsoft ITA.

Table 7, presents the estimated Logistic Regression (Logit) models for predicting the event that a tagger has used gender tags correctly, with respect to the depicted person's reported gender, using 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). We find that Black individuals are significantly less likely to receive correct gender inferences by Amazon and Clarifai on their images. Clarifai's result is similar to that of 2018, where we had found that Clarifai was less likely to infer the correct gender for Black individuals.

Abstract Supercluster The 'Abstract' tags include those indicating a subjective judgment on the person (especially about physical attractiveness), emotion/mood or state of the person, the person's personality traits, and the person's occupation or social role. Note that none of these attributes has direct visual evidence in the image, as the subjects are all depicted in a gray t-shirt with neutral facial expressions. We

	Tags	Prop. Images	Mean/ Median
C	beautiful, cute, fine looking, glamour, pretty, sexy	0.99	0.11/0.10
I	attractive, cute, handsome, pretty, sexy, smasher	0.94	0.13/0.13

Table 8: Judgment tags and frequency of use in 2021.

	Gender	Race	G*R	Sig. diff
Clarifai	710.8^{***} (0.52)	19.8^{***} (0.04)	5.8^{***} (.01)	G: W>M R: A,L,W>B
Imagga	33.2^{***} (0.05)	2.6^{*} (0.01)	7.6^{***} (0.03)	G: W>M R: A>W

Table 9: ANOVA on the use of Judgment tags in 2021.

analyze the use of these relatively few tags (used by only four of the five ITAs) with respect to the gender and race of the depicted person, remarking on any differences observed from the 2018 use and any 'new' or 'discarded' tags in 2021.

Judgments Subcluster. As listed in Table 2, only Clarifai and Imagga make use of Judgment tags in 2021, Google having discarded its only Judgment tag. Table 8 lists the tags in this subcluster for the two ITAs, along with the proportion of images on which at least one Judgment tag is used, and the mean/median number of Judgment tags used per image.

Reflecting on the 2018 results, we again observe that Clarifai and Imagga both use these types of tags frequently, although this time the roles are switched: it is *Clarifai* that uses a Judgment tag on almost every image (with Imagga following close behind). However, the number of unique tags they have are equal. We conduct an ANOVA, with gender, race, and their interaction as factors, to investigate the extent to which the use of the Judgment tags is related to the depicted person's gender and race. Table 9 presents the relevant F statistics and significance levels. For significant effects, η^2 (in parentheses) is reported, as an effect size. For each ANOVA, a Tukey post-hoc test was conducted as well. We can see that women are significantly more likely to receive Judgment tags from both ITAs. In addition, Black individuals are much less likely to receive such tags from Clarifai on their images than people of other races, while Asian individuals are more likely to receive such tags from Imagga on their images than are White individuals. These findings are completely aligned with those we detailed in 2018.

Emotion Subcluster. While now four of our five ITAs use Emotion tags, three ITAs only have one or two tags for the subcluster, and use these tags sparingly. Clarifai, with a vocabulary of five Emotion tags, uses one on most images (91%). In our previous experiments, Clarifai had used Emotion tags on only 76% of images.

Traits Subcluster. While Imagga does have one Trait tag, similar to 2018, it does not use the tag very often. Therefore, we investigated Clarifai's tags in this subcluster and found that at least one tag out of the vocabulary of 17 is used on almost every image (99%). This mirrors our 2018 findings.

	Tags	Prop. Images	Mean/ Median
Emotion			
Amazon	smile	0.15	0.012/0
Clarifai	angry, cheerful, energetic, happiness, joy, relaxation, serious, smile, surprise	0.91	0.06/0.06
Google	happy, smile	0.025	0.0031/0
Imagga	happy, smile	0.19	0.028/0
Traits			
Clarifai	attitude, casual, charming, cheerful, contemporary, cool, crazy, elegant, energetic, friendly, fun, funny, intelligence, masculinity, pensive, serious, trendy	0.99	0.26/0.26
Imagga	casual	0.025	0.002/0
Occupation			
Amazon	performer	0.011	0.001/0
Clarifai	model, scholar, son	0.17	0.009/0
Imagga	cover girl, model, representation	0.38	0.034/0

Table 10: Tags in the Emotions, Traits, and Occupations subclusters & frequency of use in 2021.

Occupation Subcluster. In 2018, only Clarifai was observed using tags from the Occupation subcluster in noteworthy quantities. There is a slight increase in the proportion of images that received Occupation tags from Clarifai, despite the reduced vocabulary. Amazon, despite adding one Occupation tag, uses it very rarely. Imagga, however, greatly increased its use of such tags in 2021.

Discussion

Between 2018 and 2021, we learned a great deal concerning the social biases inherent in computer vision applications, and which social groups are affected the most. During this time, while both Google and IBM made major announcements on how they planned to change their services in response to the growing awareness surrounding social bias, other companies' intentions were less often heard. Thus, we replicated our 2018 audit on the five ITAs still in service in 2021, to document any changing "social behaviors." While two of the ITAs did not make many substantial changes (Imagga and Microsoft), three of the five ITAs we investigated changed their service in some important ways. Not only did we see new tags for 2021, we also found that some tags from 2018 no longer appear for the same images.

Only one of the ITAs, Google, officially announced a specific change to its tags. Following the findings of research on the real-world impact of automated gender inferences (e.g. Scheuerman, Paul, and Brubaker 2019), Google removed gender-related tags from their model and it was claimed that they would "instead [..] tag any images of people with "non-gendered" labels such as "person"[18]; our results confirm that no gendered (or age-related) tags are used by Google; however, none of our images received the tag *"person"* in 2021, despite receiving it in 2018 (Barlas et al. 2021b) as well as Google's claim.

Imagga and Microsoft made no changes to their demographic tags, and their gender inference performance is almost identical to that of 2018, suggesting that the model has not changed much regarding gender or age. Microsoft's gender inferences have remained largely the same, confirming this suggestion. However, Imagga's performance has slightly improved for men's images. Imagga's performance on inferring women's gender from images remains very low.

Clarifai, the ITA with the only race/ethnicity-related tag, made no changes to its vocabulary of gender-related tags, but did increase its rate of gender inferences overall, especially inferring there to be more women in the images than in 2018. This could be seen as an attempt to "balance" the gender inference performance, especially as our audits confirm that Clarifai has improved its gender inference performance on women's images, while remaining approximately the same for men's images.

Amazon, on the other hand, was the only ITA that added demographic tags. Considering that Amazon previously had four feminine tags but only one masculine tag, the addition of one more masculine tag could also be seen as an attempt to "balance" gender inferences. Potentially as a result of this, Amazon greatly reduced the number of feminine inferences and made much more masculine inferences in 2021. Consequently, Amazon's accuracy on inferring the gender of women from their images has decreased, while its gender inferences on images of men have improved greatly.

The gender inferences are still correlated to the race of the person, as well. Black individuals have lower probability of receiving the correct gender inference from Amazon and Clarifai, as observed earlier in 2018 for Clarifai, and in other works such as (Buolamwini and Gebru 2018) with facial analysis algorithms.

Microsoft remained a fairly "conservative" ITA with no Abstract tags. Amazon, previously with no Abstract tags, added two such tags ("smile" and "performer"). Although the use of these new tags is still fairly rare, it shows a small shift in how Amazon describes people images. Google shifted its behavior slightly as well, but in the other direction: in addition to demographic tags, it also removed the only Judgment tag it was using in 2018 ("beauty").

Clarifai remains the ITA with the most inferential descriptions and vocabulary, despite a small *decrease* in the total number of unique tags, towering over the Abstract vocabularies of the other ITAs. The majority of the inferential descriptions come from the Trait subcluster, which Clarifai used to describe almost every image in 2021, as in 2018. Similarly, Clarifai uses at least one Judgment tag on almost every image as well, although this is an increase from 2018. Imagga, although with a smaller vocabulary of Abstract tags, goes head-to-head with Clarifai in how often it uses its Judgment tags. With both ITAs in 2021, women are more likely to receive Judgment tags on their images, and some races (Black individuals with Clarifai, Asian individuals with Imagga) are less likely to receive such tags.

[18]https://www.businessinsider.com/google-cloud-vision-api-wont-tag-images-by-gender-2020-2

Amazon and Google used new tags relating to observable, concrete concepts in the image that were not observed in 2018. Most additions were in the Clothing subcluster, potentially showing a larger focus on fashion. Some tags describing hairstyles such as "afro_hairstyle" and "mohawk" were removed from Amazon, potentially because they are (perceived to be) connected to a particular racial culture.[19,20,21] At the same time, we observed a new tag in the Google outputs, "cornrows", which also has a racialized nature (e.g. Caldwell 1991, on braided hairstyles). So while the ITAs may be shifting their behavior and vocabulary in one area (gender or racialized hairstyles) due to the changing social awareness, they may not be making progress in other areas.

Having seen a wide range of changes – from removing every tag referring to gender and age, to removing just some color tags – it is difficult to imagine how developers and other users keep up with the services' social behaviors. The update emails and announcements rarely mention specific changes to the model, especially when the changes are not adding/removing tags but instead modifying the tag's concept within the model such that it is predicted for different images. In those cases, even if the users had tested the ITA and set up the pipeline accordingly at the beginning of their relationship with the service, there is no guarantee that everything will run the same way one or two years later. On top of that, since temporal audits of such services are almost non-existent, the differences will not even be noticed until they have affected many outputs.

Perhaps the changes made to these systems are overlooked as the tags are assumed to be "harmless" and objective, and it is probably true that the false negative of a tag such as "green" will not break a system or cause harm to people in the images. However, there are tags which make inferences about the depicted person's personality, mood, or physical attractiveness; and we have found that such tags are not used in the same manner regarding the person's race and gender. In addition, the inferences about a person's demographics, especially when such inferences cannot be made from a person's appearance in the first place, can have a negative impact. Therefore, users of these systems as well as developers who use these services in their products, really need to be aware of the changes to the services, especially regarding these types of tags.

Cognitive services such as ITAs allow all kinds of users access to state-of-the-art computer vision tools. Users may only know the basics of how these services work, and may not care to learn further beyond connecting the service to their application pipeline. Therefore, expecting audits to happen at the user end is not realistic and not a good placement of the accountability. The service providers should be responsible for communicating any changes to the model to the public and their users; in addition to that, external (third-party) auditors should be required to test the algorithms to ensure the outputs maximize public good, while minimizing potentially harmful (use of) tags.

Recently, two new datasets have been released by the Chicago Face Database team, depicting i) multiracial and ii) Indian individuals in the same manner as the original dataset. The Multiracial dataset depicts 88 individuals "who self-reported multiracial ancestry" recruited in the U.S. (Ma, Kantner, and Wittenbrink 2021), while the India dataset depicts 142 individuals recruited in Delhi, India (Lakshmi et al. 2021). Our current RQs pertain to the temporal changes in the ITA models; therefore, we do not analyze these two datasets or the outputs we receive for them from the ITAs in this paper. However, it is interesting to note this development, as it appears it is not only the service providers who are moving towards a more socially-aware practice, but also researchers and dataset creators. Future work may use these datasets (along with other datasets with diverse identities) to see further the behaviors of the ITAs, especially with respect to other racial and ethnic groups.

Limitations

Replication studies are a powerful tool for revealing change over time. However, the reasons for those changes cannot be uncovered by replication alone. As such, any suggestions we make as to *why* we observe differences in our results are merely hypotheses, informed by the state of the field.

As with the 2018 study, we have only used standardized portrait images of people. While this highly controlled set of inputs allows us to investigate our research questions effectively, it is far removed from the real world use cases, in which images of all manners are used. In these images, surroundings and background context might influence the output (e.g. Barlas et al. 2021a).

As mentioned in an earlier section, the fact that a certain tag did not appear in our results in 2021 does not mean that the tag was removed entirely from that service. It is possible that certain tags also existed in the vocabulary of the service in 2018, and continue to exist now, but for some reason (e.g., a change in the model) did not receive a high enough confidence score to appear for our CFD images' tags.

Conclusion

Transparency on the service providers' end is essential to understanding and keeping up with these tools. Without transparency, neither the users of the services nor third parties affected by the services can see the effects of the outputs. On top of that, the lack of transparency regarding changes over time means that harmful patterns in the outputs of these services may take longer to notice or identify, having more negative effects by the time audits uncover the problems.

On both sides of the Atlantic, we are still at the point where algorithmic services such as ITAs are not regulated, although this is likely to change in the near future, particularly within the European Union. In the meantime, we must find other ways to discover and rectify the harmful effects. Audits of social behaviors of algorithms, coupled with temporal audits such as ours (i.e., conducted periodically), will be necessary to have some oversight.

While some changes observed in the ITAs are in a beneficial direction – like that of Google's, following research and

[19] booksandideas.net/The-Afro-More-Than-a-Hairstyle.html

[20] embracerace.org/resources/why-we-dont-wear-mohawks

[21] melmagazine.com/en-us/story/a-spiky-history-of-the-mohawk

real-world impacts of the services – others may be harmful. However, without oversight mechanisms in place, it is impossible to tell. Our audits combine manual, qualitative analysis with quantitative analyses; however, to truly scale up these audits, future work should investigate whether such audits can be automated, and if so, whether the results of the audits change periodically.

Acknowledgments

This project has received funding from the Cyprus Research and Innovation Foundation under grant EXCELLENCE/0918/0086 (DESCANT), the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreements No. 739578 (RISE) & No. 810105 (CyCAT), and the Government of the Republic of Cyprus through the Deputy Ministry of Research, Innovation and Digital Policy.

References

Asplund, J.; Eslami, M.; Sundaram, H.; Sandvig, C.; and Karahalios, K. 2020. Auditing race and gender discrimination in online housing markets. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, 24–35.

Bandy, J. 2021. Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1).

Barlas, P.; Kyriakou, K.; Guest, O.; Kleanthous, S.; and Otterbacher, J. 2021a. To "See" is to Stereotype: Image Tagging Algorithms, Gender Recognition, and the Accuracy-Fairness Trade-Off. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW3).

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

Barlas, P.; Kyriakou, K.; Kleanthous, S.; and Otterbacher, J. 2021b. *Person, Human, Neither: The Dehumanization Potential of Automated Image Tagging*, 357–367. New York, NY, USA: Association for Computing Machinery. ISBN 9781450384735.

Buolamwini, J.; and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, 77–91. PMLR.

Caldwell, P. M. 1991. A hair piece: Perspectives on the intersection of race and gender. *Duke Law Journal*, 365–96.

Chen, I.; Johansson, F. D.; and Sontag, D. 2018. Why is my classifier discriminatory? *arXiv preprint:1805.12002*.

Christoforou, E.; Barlas, P.; and Otterbacher, J. 2021. It's About Time: A View of Crowdsourced Data Before and During the Pandemic. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–14.

Hornbæk, K.; Sander, S. S.; Bargas-Avila, J. A.; and Grue Simonsen, J. 2014. Is once enough? On the extent and content of replications in human-computer interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 3523–3532.

Jung, S.-G.; An, J.; Kwak, H.; Salminen, J.; and Jansen, B. J. 2018. Assessing the accuracy of four popular face recognition tools for inferring gender, age, and race. In *Twelfth international AAAI conference on web and social media*.

Karkkainen, K.; and Joo, J. 2021. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1548–1558.

Kay, M.; Matuszek, C.; and Munson, S. A. 2015. *Unequal Representation and Gender Stereotypes in Image Search Results for Occupations*, 3819–3828. New York, NY, USA: Association for Computing Machinery. ISBN 9781450331456.

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

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

Kyriakou, K.; Kleanthous, S.; Otterbacher, J.; and Papadopoulos, G. A. 2020. Emotion-Based Stereotypes in Image Analysis Services. In *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '20 Adjunct, 252–259. New York, NY, USA: Association for Computing Machinery. ISBN 9781450379502.

Lakshmi, A.; Wittenbrink, B.; Correll, J.; and Ma, D. S. 2021. The India Face Set: International and Cultural Boundaries Impact Face Impressions and Perceptions of Category Membership. *Frontiers in psychology*, 12: 161.

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.

Ma, D. S.; Kantner, J.; and Wittenbrink, B. 2021. Chicago Face Database: Multiracial expansion. *Behavior Research Methods*, 53(3): 1289–1300.

Mashhadi, A.; Winder, S. G.; Lia, E. H.; and Wood, S. A. 2021. No Walk in the Park: The Viability and Fairness of Social Media Analysis for Parks and Recreational Policy Making. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, 409–420.

Metaxa, D.; Gan, M. A.; Goh, S.; Hancock, J.; and Landay, J. A. 2021. An Image of Society: Gender and Racial Representation and Impact in Image Search Results for Occupations. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1): 1–23.

Otterbacher, J.; Bates, J.; and Clough, P. 2017. Competent men and warm women: Gender stereotypes and backlash in image search results. In *Proceedings of the 2017 chi conference on human factors in computing systems*, 6620–6631.

Phillips, P. J.; Jiang, F.; Narvekar, A.; Ayyad, J.; and O'Toole, A. J. 2011. An Other-Race Effect for Face Recognition Algorithms. *ACM Trans. Appl. Percept.*, 8(2).

Rahwan, I. 2018. Society-in-the-loop: programming the algorithmic social contract. *Ethics and Information Technology*, 20(1): 5–14.

Raji, I. D.; and Buolamwini, J. 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 429–435.

Salton, G.; and Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5): 513–523.

Scheuerman, M. K.; Paul, J. M.; and Brubaker, J. R. 2019. How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis Services. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).

Taddeo, M.; and Floridi, L. 2018. How AI can be a force for good. *Science*, 361(6404): 751–752.

Yang, K.; Qinami, K.; Fei-Fei, L.; Deng, J.; and Russakovsky, O. 2020. Towards Fairer Datasets: Filtering and Balancing the Distribution of the People Subtree in the ImageNet Hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, 547–558. New York, NY, USA: Association for Computing Machinery. ISBN 9781450369367.

Ye, T.; You, S.; and Robert Jr, L. 2017. When does more money work? Examining the role of perceived fairness in pay on the performance quality of crowdworkers. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

Zeng, Z.; Islam, R.; Keya, K. N.; Foulds, J.; Song, Y.; and Pan, S. 2021. Fair Representation Learning for Heterogeneous Information Networks. *Proceedings of the International AAAI Conference on Web and Social Media*, 15(1): 877–887.