Person, Human, Neither: The Dehumanization Potential of Automated Image Tagging

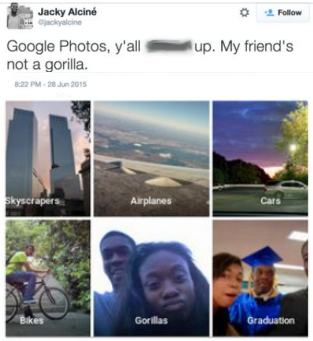
Pınar Barlas,[1] Kyriakos Kyriakou,[1] Styliani Kleanthous,[1,2] & Jahna Otterbacher[1,2]
[1] CYENS Centre of Excellence
[2] Cyprus Center for Algorithmic Transparency

Our Goal.

We investigate six "image tagging algorithms" (ITAs) for their potential to dehumanize.

Motivation.

Many algorithms embed and reinforce harmful social biases in their outputs; one example from 2015 demonstrates an explicit example of **dehumanization**, where the Google Photos app automatically tagged images of two Black people as 'gorillas'.



Research Questions & Method.

RQ1: What **vocabulary** do the image tagging algorithms use to **identify humans** ('humanness tags')? What kind of dehumanizing effect may an **error** have?

The denial of humanness or human qualities to a person, through an explicit declaration, is termed *blatant, metaphor-based, mechanistic dehumanization* in the social psychology literature. We explore the way (mis)use of humanness tags may replicate this behavior in an automated manner.

We manually select the tags which indicate the presence of a person in the image ("humanness tags") from each ITA. Where an ITA uses more than one such tag, we compare the use of the tags.

RQ2: Do the ITAs apply this vocabulary to **every photo** in a controlled, diverse dataset of people images? Do **errors** appear for any **social groups** more than others?

Following similar audits of image analysis algorithms, we examine the outputs for a controlled set of inputs. In order to minimize the variables in the image, we use the Chicago Face Dataset as our inputs: a set of standardized portrait photographs of people, all wearing a grey t-shirt and with a neutral expression.

The outputs are grouped into the depicted person's race and gender and examined for dehumanizing treatment (looking at 'group fairness').

RQ3: Do the ITAs identify **faces**? If so, are there any photos where the **face is identified** but the **'human' isn't**?

Aware of the dehumanizing behaviors which are enabled by seemingly harmless use of technology, we question whether this technology is more readily available to enable surveillance than it is to detect the 'humanness' of the people in the images.

For this purpose, we collect the tags which indicate the presence of a face in the image, and compare the frequency of this tag to those of the 'humanness' tags examined earlier. We investigate whether there are images which received a 'face' tag but not a 'humanness' tag, or whether there are those which received neither type of tag.

Findings.

- All six ITAs have at least one tag to identify a human in the image [Table 1].
- All ITAs used the "person" tag, except for Clarifai which used "no person" instead [Table 1].
- Two ITAs used the "human" tag as well; Amazon used both tags on every image, while Google did not tag any image with both.
- Watson had three textually overlapping tags, but the co-occurrence analysis [Figure 1] indicates the three tags are used in distinct manners. No image received all three tags.

- Clarifai's "no person" tag was used twice, and these two images are included in our discussions of "images which did not receive humanness tags", or *H'*.
- Although three tags (Amazon's "person" and "human; Microsoft's "person") were used on every image on the dataset, the majority of the tag x ITA combinations had some margin of error [Table 1].
- The remaining four ITAs had at least one image which did not have any 'humanness' tags, ranging from one (Watson) to 492 (Google) images in our total 597 [Table 1]. Higher proportions of women and of Black people had their images receive no humanness tags, with Black Women's images receiving the highest error rates [Table 2].

- All six ITAs have a 'face' tag, and four ITAs correctly identify the face in the image most of the time, even though ITAs are not specialized in facial recognition or analysis [Table 1].
- Images that received no humanness tags often still received a 'face' tag [Table 2]. Particularly, the two images Clarifai tagged with 'no person' both also received the 'face' tag; the people depicted in the two images were both Black.

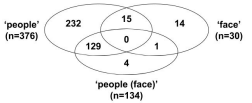


Figure 1: Co-occurrence analysis of Watson's three tags.

Figure 2: Venn Diagram demonstrating *H'* (the shaded area), the set of images which did not receive any humanness tags.

	A	C	G	I	M	W
'person'	597	'no person': 2	85	589	597	596
'human'	597	-	20	-	-	-
'people'	596	587	-	45	-	376
'people (face)'	-	-	-	-	-	134
'face' *(F)*	597	466	586	597	57	30
No humanness tags *(H')*	0	12*	492	8	0	1
Lacking humanness & face tags *(H'∩F')*	0	6	10	0	0	1

Table 1: Frequencies of the "humanness tags", the 'face' tag, and the number of images lacking humanness and/or face tags, for each ITA. *Includes the two images with the 'no person' tag.

	n	Clarifai	Google	Imagga	Watson
Asian Women	57	1/1	56/56	-	-
Asian Men	52	-	40/40	-	-
Black Women	104	1/2*	82/91	5/5	0/1
Black Men	93	1/1*	67/68	-	-
Latina Women	56	0/1	56/56	2/2	-
Latino Men	52	0/1	31/31	-	-
White Women	90	2/5	86/86	1/1	-
White Men	93	1/1	64/64	-	-
Total	**597**	**6/12***	**482/492**	**8/8**	**0/1**

Table 2: Each cell: *n* of *(H'∩F) /H'* for a particular ITA and social group. *Includes two images which received both the 'no person' and the 'face' tags.

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreements No 739578 and No 810105.

This project has received funding from the Government of the Republic of Cyprus through the Deputy Ministry of Research, Innovation and Digital Policy.

© CYENS Centre of Excellence