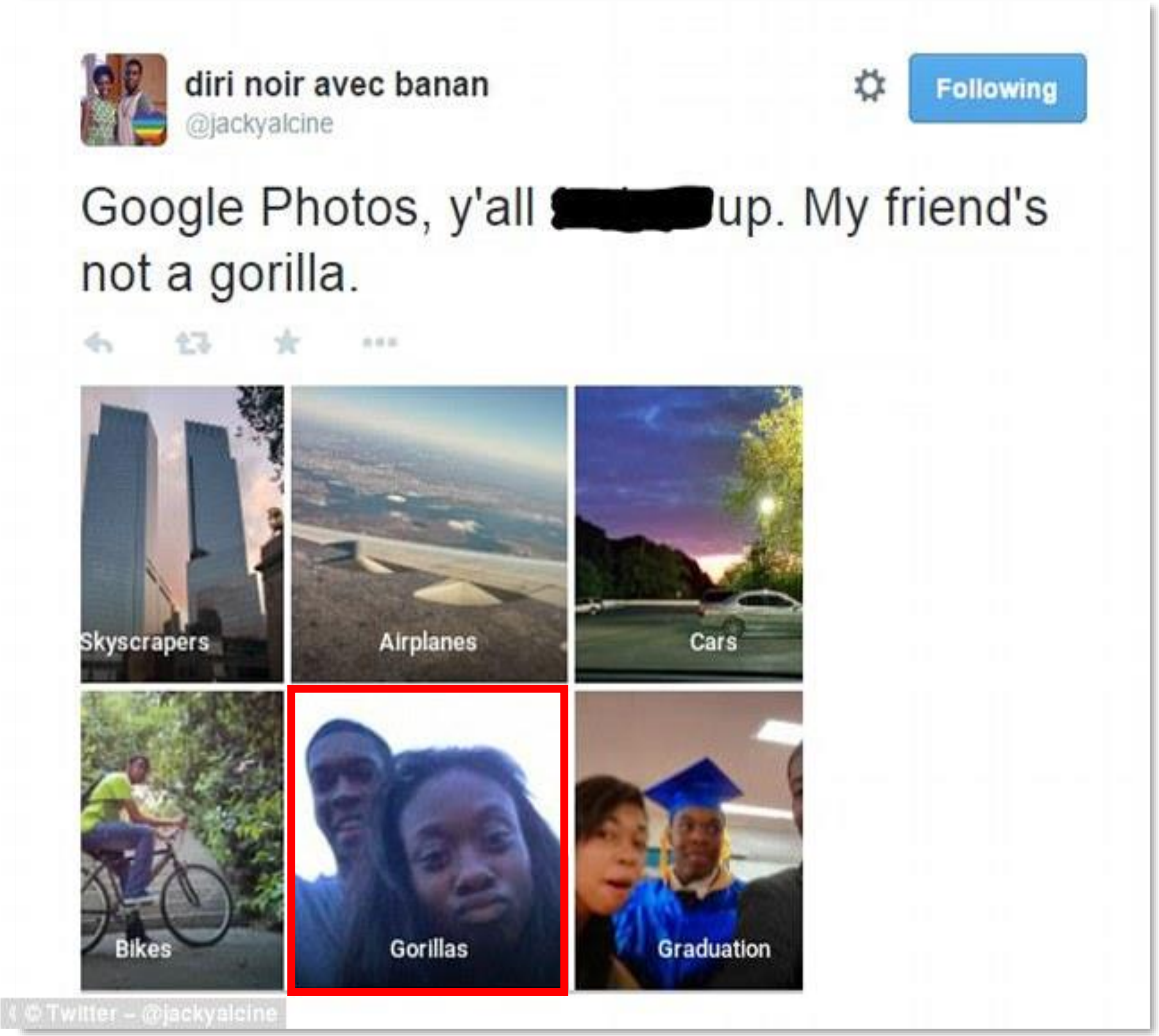
Measuring Embedded Human-like Biases in Face Recognition Models

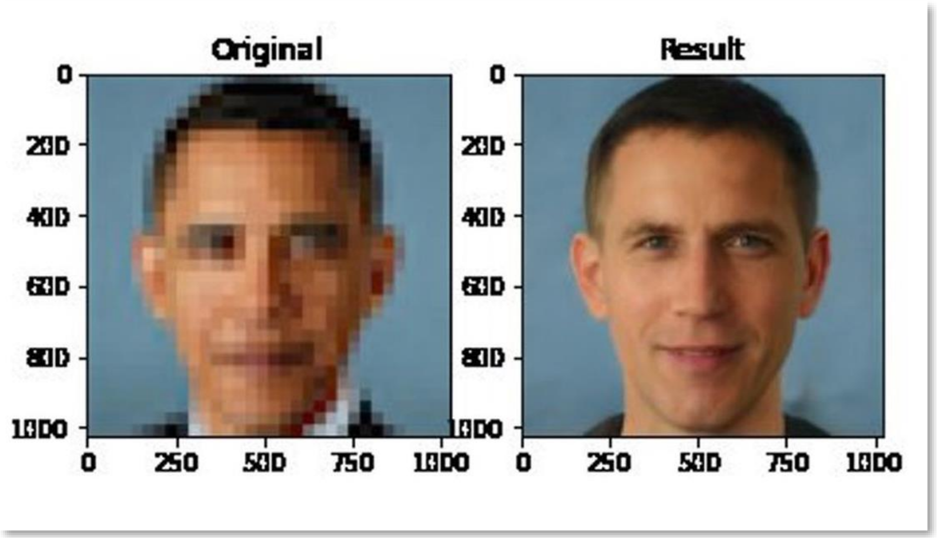
SangEun Lee*　Soyoung Oh*　Minji Kim　Eunil Park

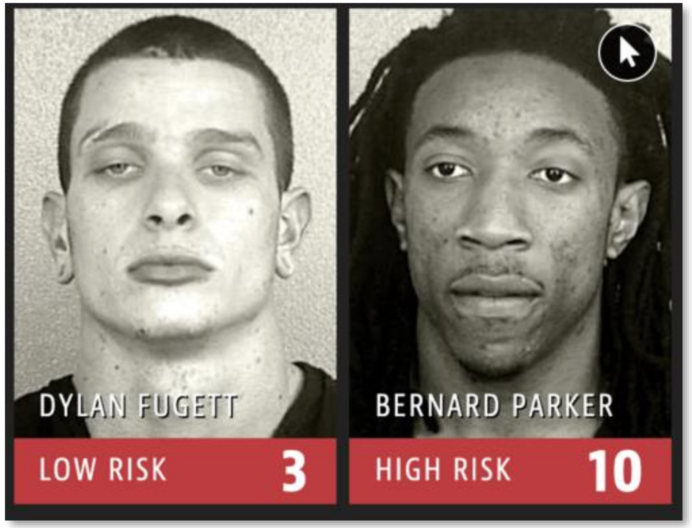
Department of Applied Artificial Intelligence, Sungkyunkwan University(SKKU)

{sangee1104, sori424, m5512m}@g.skku.edu, eunilpark@skku.edu

Issues with Face Recognition Models







Face Embedding Association Test (FEAT) (1/2)

FEAT measures social bias in the face recognition models by comparing the relative association between targets and attributes.

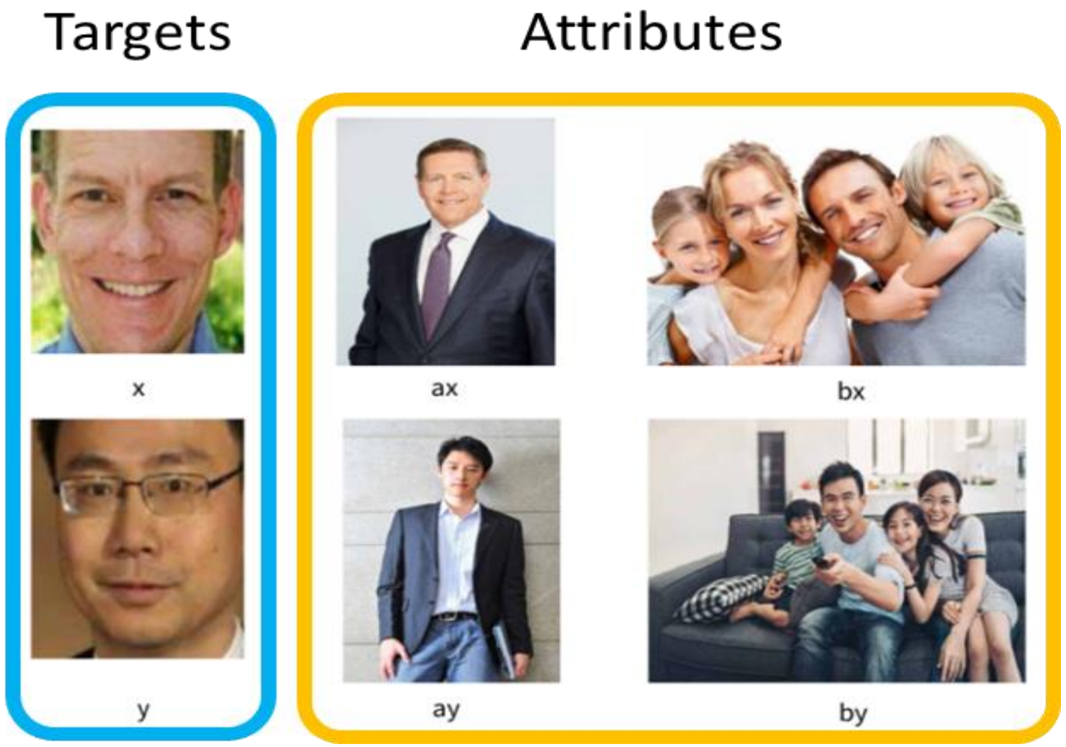


Figure 1: An example of target and attribute sets

Face Embedding Association Test (FEAT) (2/2)

FEAT measures **the relative association** between two sets of target concepts and two sets of attributes

Target Image Sets

X = { , …   } ≈ European American

Y = { , …   } ≈ Asian American

Attribute Image Sets

A = { , , …   } ≈ Career

B = { , , …   } ≈ Family

$$s(f, A, B) = [mean_{a \in A} \cos(f, a) - \ mean_{b \in B} \cos(f, b)]$$

$$Effect\ size = \frac{mean_{x \in X}\, s(x, A, B) - \ mean_{y \in Y}\, s(y, A, B)}{std\_dev_{f \in X \cup Y}\, s(f, A, B)}$$

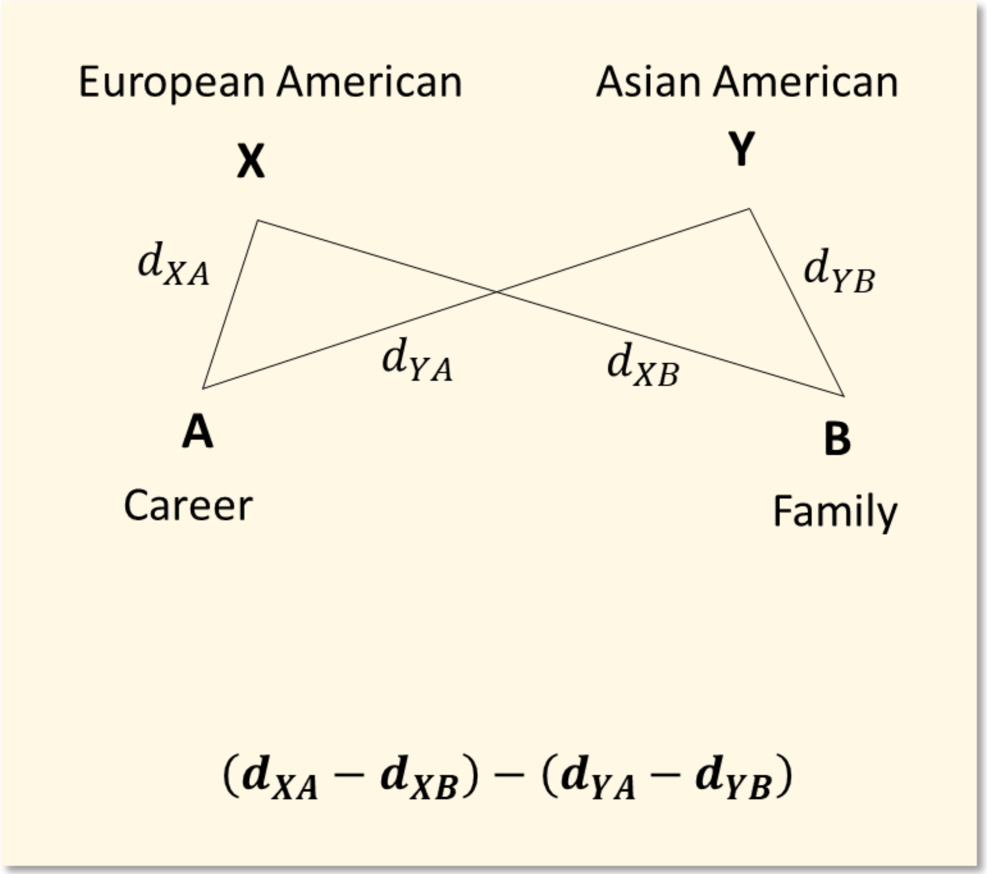


Figure 2: An example of face embedding association test

Research Questions

1) Do face recognition models contain racial bias?
2) Do face recognition models contain gender bias?
3) Do face recognition models contain age bias?
4) Do face recognition models contain intersectional bias?

Corresponding Concepts with RQs (1/2)

Target Sets

1) Race

European American

African American

Asian American

2) Gender

Male

Female

3) Age

Young

Old

4) Intersectional

European American Female

African American Female

Asian American Female

Corresponding Concepts with RQs (2/2)

Attribute Sets[1][2]

1) Career & Family


2) Pleasant & Unpleasant

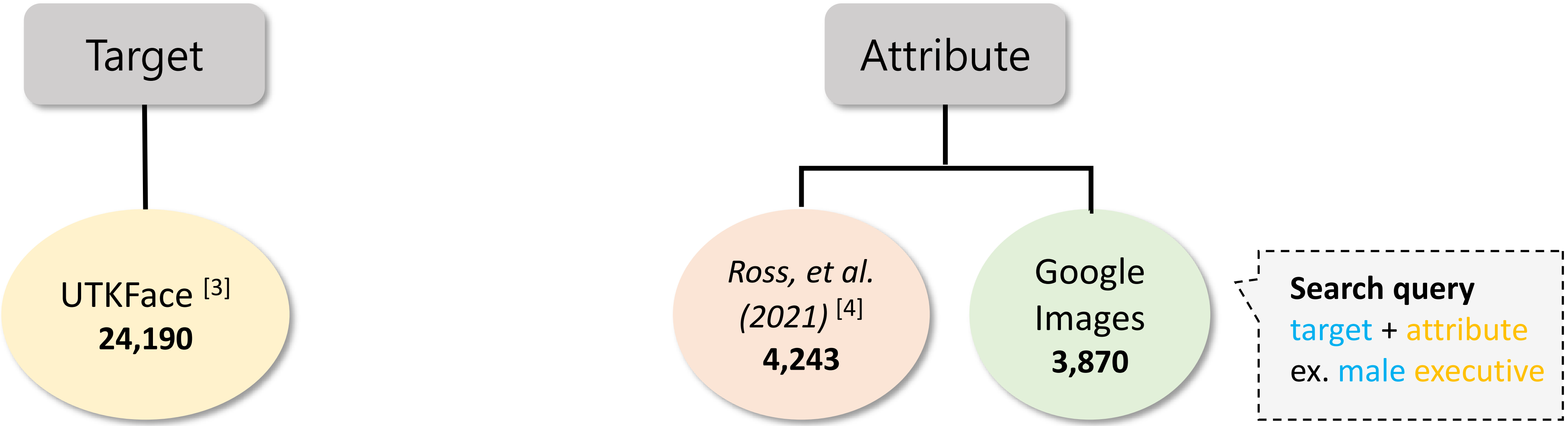

3) Likable & Unlikable


4) Competent & Incompetent


[1] Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. Journal of personality and social psychology, 74(6), 1464.
[2] Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334), 183-186.

Experimental Setup

- Data Collection



- Pre-trained Models
  - DeepFace, DeepID, VGGFace, FaceNet, OpenFace and ArcFace

[3] Zhang, Z., Song, Y., & Qi, H. (2017). Age progression/regression by conditional adversarial autoencoder. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5810-5818).
[4] Ross, C., Katz, B., & Barbu, A. (2021, June). Measuring Social Biases in Grounded Vision and Language Embeddings. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 998-1008).

1. Do face recognition models contain racial bias?

Attributes	Targets	DeepFace	DeepID	VGGFace	FaceNet	OpenFace	ArcFace
Career/Family	EA/AA	**0.095***	**0.078***	**0.294***	**0.569***	**0.148***	-0.001
	EA/AS	-0.006	-0.209	-0.476	-0.097	**0.372***	**0.078***
Pleasant/Unpleasant	EA/AA	**0.507***	**0.557***	**0.939***	**1.081***	**0.635***	**0.277***
	EA/AS	-0.049	-0.001	-0.138	0.009	0.140*	**0.165***
Likable/Unlikable	EA/AA	**0.134***	**0.647***	0.021	**1.084***	**0.287***	**0.517***
	EA/AS	-0.032	-0.112	-0.829	-0.121	**0.111***	-0.524
Competent/Incompetent	EA/AA	-0.038	-0.520	-1.215	**0.704***	-0.575	-0.200
	EA/AS	0.012	**0.075***	**0.223***	-0.123	-0.333	**0.186***

Table 1: European American (EA), African American (AA), Asian American(AS), $p<0.05$*

Yes. Effect size represents measurable biases for all models.

2. Do face recognition models contain gender bias?

Attributes	Targets	DeepFace	DeepID	VGGFace	FaceNet	OpenFace	ArcFace
Career/Family	M/F	0.002	-0.412	-0.197	-0.106	**0.445***	**0.111***
Pleasant/Unpleasant	M/F	0.001	-0.194	-0.089	-0.042	0.020	**0.452***
Likable/Unlikable	M/F	0.002	-0.053	-0.030	**0.237***	0.053	-0.243
Competent/Incompetent	M/F	-0.001	-0.036	**0.205***	-0.343	**0.212***	0.035

Table 2: Male (M), Female (F), p<0.05*

Yes. Less than racial bias, still effect size represents measurable biases toward gender for VGGFace, FaceNet, OpenFace, and ArcFace.

3. Do face recognition models contain age bias?

Attributes	Targets	DeepFace	DeepID	VGGFace	FaceNet	OpenFace	ArcFace
Career/Family	Y/O	-0.055	-0.376	**0.344***	-0.166	0.993	-0.416
Pleasant/Unpleasant	Y/O	0.062	-0.036	**1.406***	0.137	**0.551***	-0.260
Likable/Unlikable	Y/O	0.066	**0.290***	**1.222***	0.000	**0.431***	**0.509***
Competent/Incompetent	Y/O	-0.021	-0.001	**1.046***	0.031	**0.225***	-0.477

Table 3: Young (Y), Old (O), $p<0.05$*

Yes. Effect size represents measurable biases toward age group for DeepID, VGGFace, OpenFace, and ArcFace.

4. Do face recognition models contain intersectional bias?

Intersectional bias

- "Asian women are considered as incompetent; not a leader, submissive, and expected to work at a low-level gendered job."[5]

Attributes	Targets	DeepFace	DeepID	VGGFace	FaceNet	OpenFace	ArcFace
Competent/Incompetent	EAF/AAF	-0.017	**0.465***	-1.007	**0.748***	-0.095	**0.358***
	EAF/ASF	0.006	-0.172	0.029	**0.165***	-0.237	**0.354***
	AAF/ASF	0.072	0.017	**1.424***	**0.451***	**0.453***	-0.367

Table 4: European American Female (EAF), African American Female (AAF), Asian American Female (ASF), $p<0.05$*

Yes. Effect size represents a measurable bias toward intersectional groups for all models except DeepFace.

[5] Mukkamala, S., & Suyemoto, K. L. (2018). Racialized sexism/sexualized racism: A multimethod study of intersectional experiences of discrimination for Asian American women. Asian American journal of psychology, 9(1), 32.

Race sensitivity analysis

- To explore whether the racially-dependent external features result in racial bias in models.
  → Gradually reversed the racial features of images; i.e. European American ⇄ African American

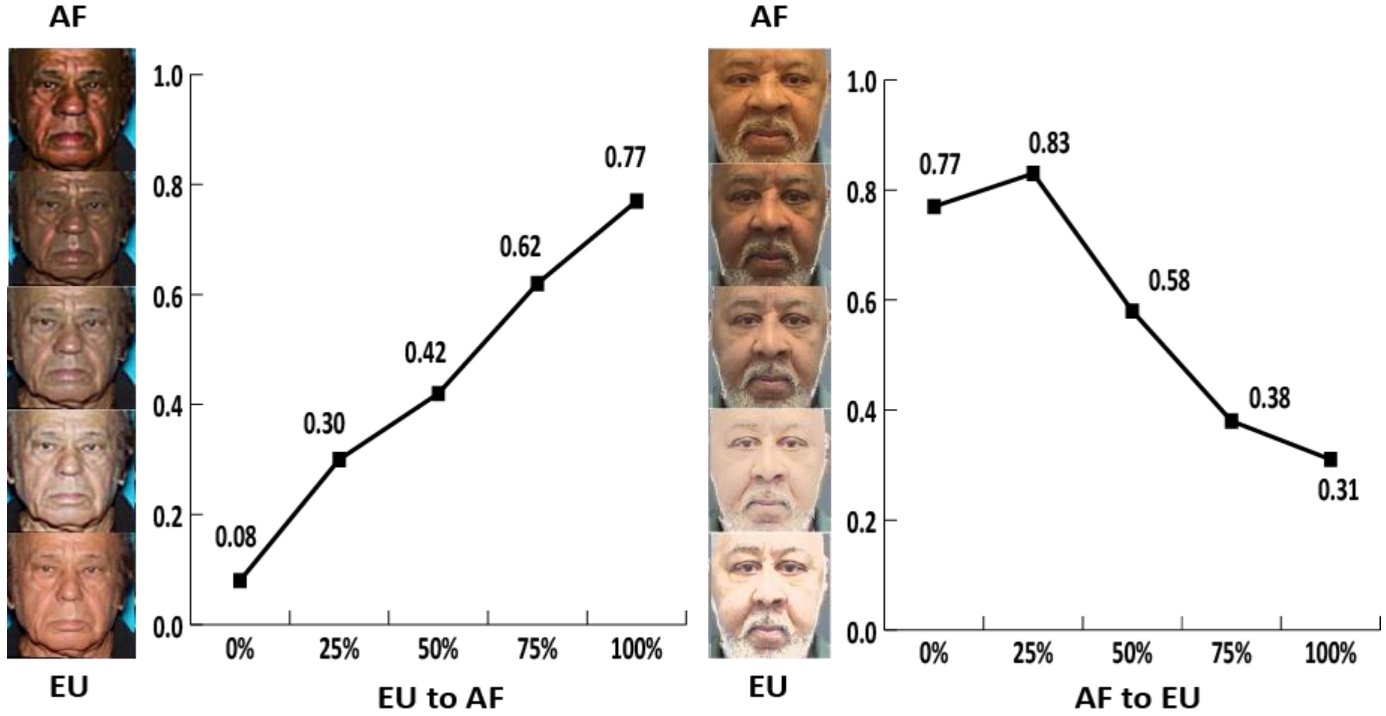


Figure 3: The classification probability of race between AF and EU by extent of the race transformation

- Before that, we validated whether a model classifies the race differently as the race of the image is converted.

Race sensitivity analysis

Race Transformation	Attributes	DeepFace	DeepID	VGGFace	FaceNet	OpenFace	ArcFace
25%	Career/Family	0.598*	0.470*	0.354*	0.419*	0.657*	0.523*
	Pleasant/Unpleasant	0.438*	0.314*	1.723*	0.720*	0.267*	0.901*
	Likable/Unlikable	0.796*	0.202*	1.414*	0.607*	0.756*	0.077
	Competent/Incompetent	0.957*	0.717*	1.420*	0.645*	1.306*	0.657*
50%	Career/Family	-0.007	-0.560	-0.689	-0.770	-0.281	-0.443
	Pleasant/Unpleasant	-0.029	-0.409	1.591*	-0.754	-0.510	0.201*
	Likable/Unlikable	0.008	-0.961	0.834*	-0.729	-0.378	-0.951
	Competent/Incompetent	-0.095	-0.624	0.817*	-0.716	0.308*	-0.501
75%	Career/Family	-0.768	-1.226	-1.362	-1.467	-1.134	-1.089
	Pleasant/Unpleasant	-0.653	-0.888	1.324*	-1.547	-1.188	-0.475
	Likable/Unlikable	-1.018	-1.515	-0.387	-1.490	-1.318	-1.375
	Competent/Incompetent	-1.170	-1.439	-0.549	-1.509	-1.036	-1.278
100%	Career/Family	-1.112	-1.538	-1.586	-1.725	-1.490	-1.382
	Pleasant/Unpleasant	-0.999	-1.200	0.761*	-1.785	-1.493	-0.884
	Likable/Unlikable	-1.448	-1.733	-1.102	-1.745	-1.619	-1.593
	Competent/Incompetent	-1.536	-1.697	-1.046	-1.755	-1.493	-1.628

Table 5: The results for race sensitivity analysis with FEAT on race transformation

Race sensitivity analysis

Race Transformation	Attributes	DeepFace	DeepID	VGGFace	FaceNet	OpenFace	ArcFace
25%	Career/Family	0.598*	0.470*	0.354*	0.419*	0.657*	0.523*
	Pleasant/Unpleasant	0.438*	0.314*	1.723*	0.720*	0.267*	0.901*
	Likable/Unlikable	0.796*	0.202*	1.414*	0.607*	0.756*	0.077
	Competent/Incompetent	0.957*	0.717*	1.420*	0.645*	1.306*	0.657*
[illegible]	Career/Family	-0.007	-0.560	-0.689	-0.770	-0.281	-0.443
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
75%	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	Pleasant/Unpleasant	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	Likable/Unlikable	-1.018	-1.515	-0.387	-1.490	-1.318	-1.375
	Competent/Incompetent	-1.170	-1.439	-0.549	-1.509	-1.036	-1.278
100%	Career/Family	-1.112	-1.538	-1.586	-1.725	-1.490	-1.382
	Pleasant/Unpleasant	-0.999	-1.200	0.761*	-1.785	-1.493	-0.884
	Likable/Unlikable	-1.448	-1.733	-1.102	-1.745	-1.619	-1.593
	Competent/Incompetent	-1.536	-1.697	-1.046	-1.755	-1.493	-1.628

External racial features can be the cause of discriminative associations in the embedding space.

Table 5: The results for race sensitivity analysis with FEAT on race transformation

Discussion

What we have done…

- Investigated 6 face recognition models across 4 biases.
- Confirmed racial, gender, age, and an intersectional bias are reproduced through the embeddings from pre-trained models.
- Suggested a wide range of subgroup and ethnicity should be considered with respect to examining social biases.

What are next steps?

- Identify the source of reproducing bias, data distribution or algorithmic bias.
- Bias mitigation techniques would be presented.

Data eXperience Lab

SUNG KYUN KWAN UNIVERSITY(SKKU)

Measuring Embedded Human-like Biases in Face Recognition Models

Feel free to reach out!

SangEun Lee (sange1104@g.skku.edu)

Thank you ☺

Our data and code is publicly released!