Responsible AI: Measuring accuracy and bias in Computer Vision Systems

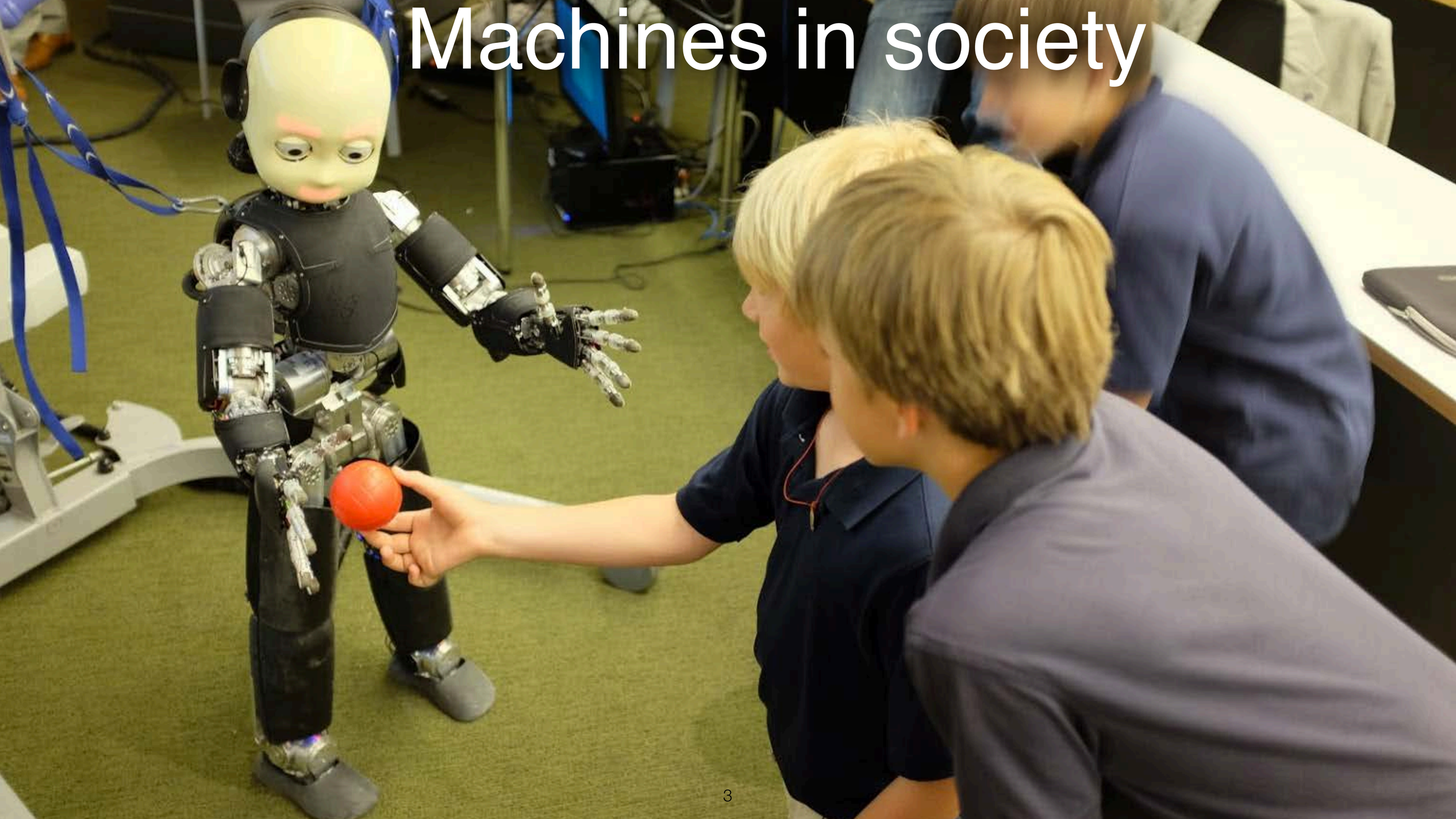
Pietro Perona
Caltech and AWS

Ellis Summer School on Large Scale AI
Modena - 18-22 September 2023

Plan

- Responsible AI
- Bias and fairness
- Correlation vs causation
- Bias in face analysis and recognition
- Bias in generative models
- Fixing bias
- Conclusions

Machines in society



Responsible AI

- Fairness, bias

- Robustness
- Explainability
- Safety
- Privacy
- IP

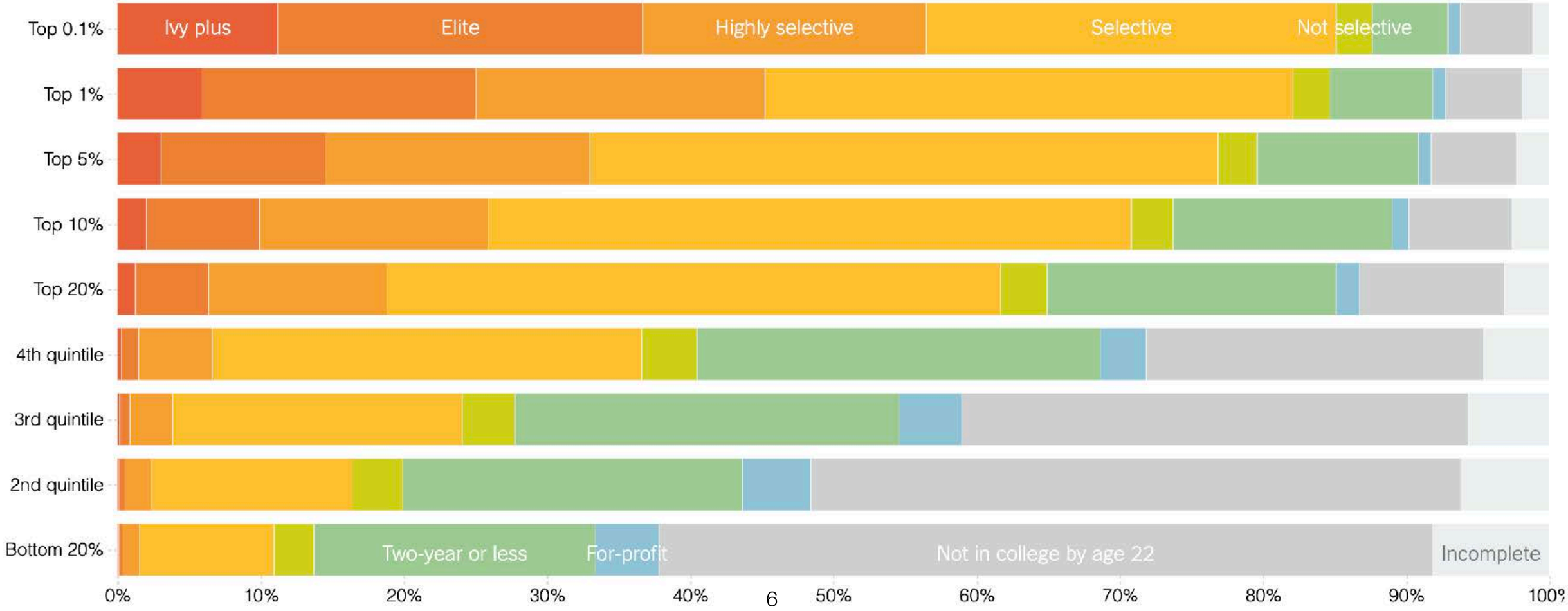
WORLD & NATION

Identical twins. Identical asylum claims. Very different luck at the border

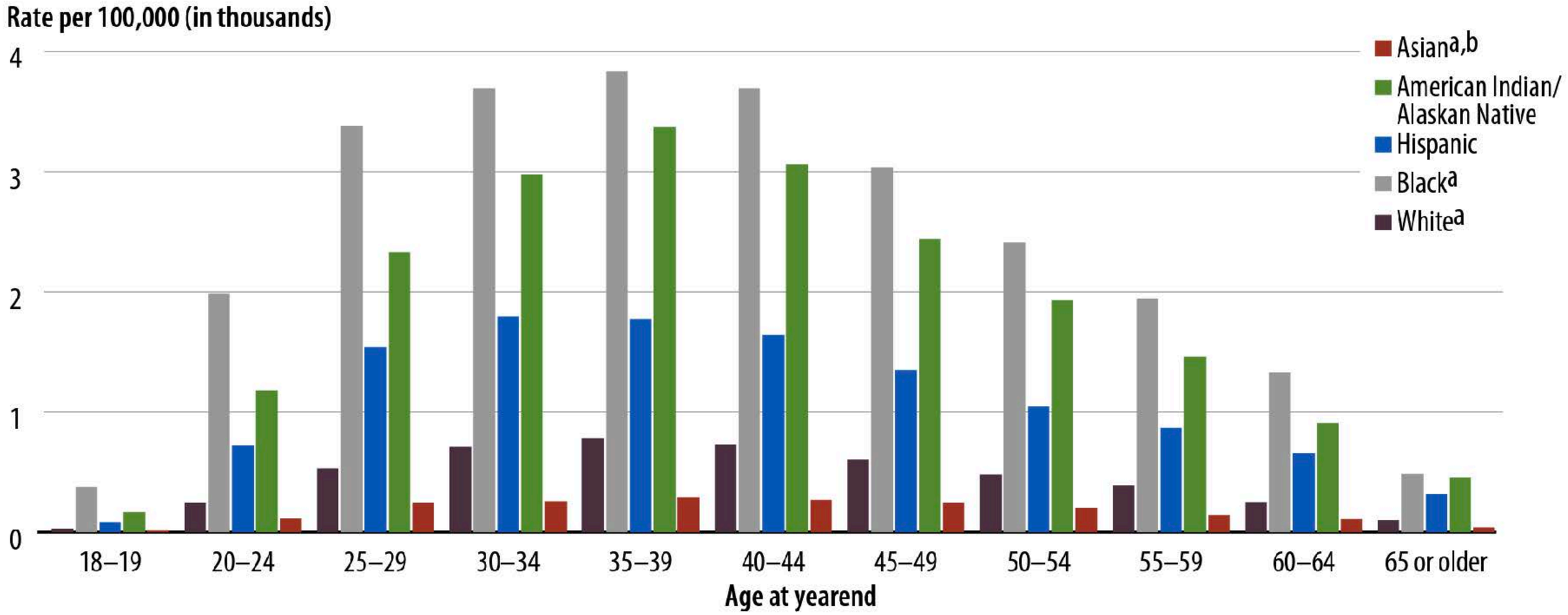
Where today's 25-year-olds went to college, grouped by their parents' income

The New York Times

About four in 10 students from the top 0.1 percent attend an Ivy League or elite university, roughly equivalent to the share of students from poor families who attend any two- or four-year college.



Imprisonment rates of male U.S. residents, based on sentenced prisoners under the jurisdiction of state or federal correctional authorities, by demographic characteristics, December 31, 2021



Males: 1,084,717
Females: 78,948

Fairness

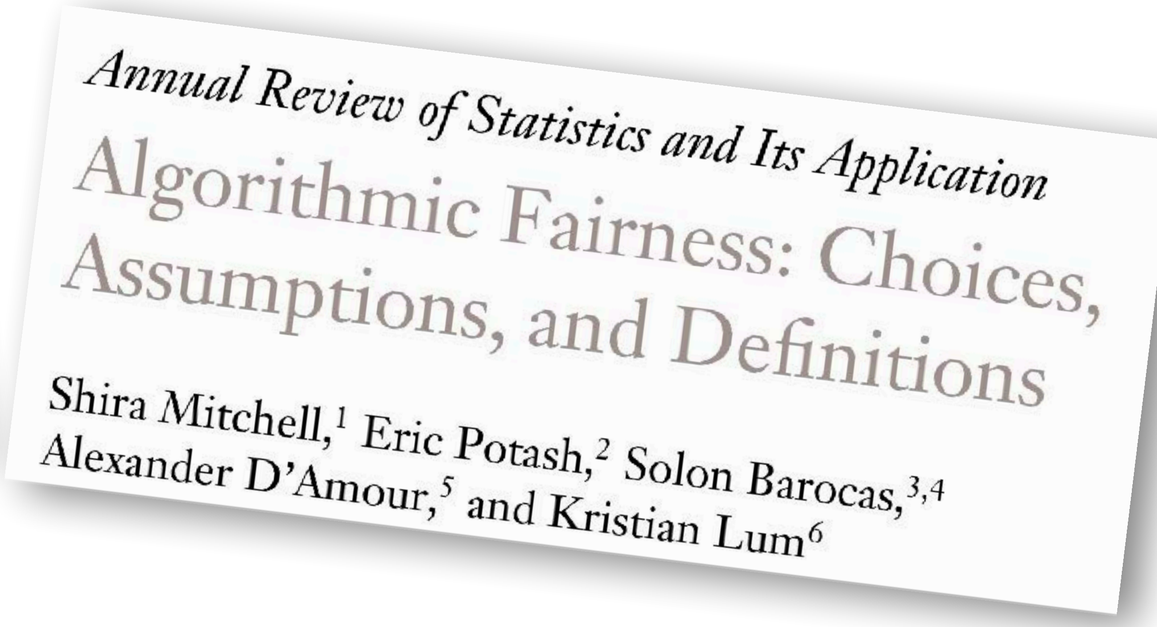
Input (e.g. action, money) → f(.) → *Outcomes*


AEA Papers and Proceedings 2018, 108: 22–27
https://doi.org/10.1257/pandp.20181018

ADVANCES IN BIG DATA RESEARCH IN ECONOMICS

Algorithmic Fairness†

By Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan*

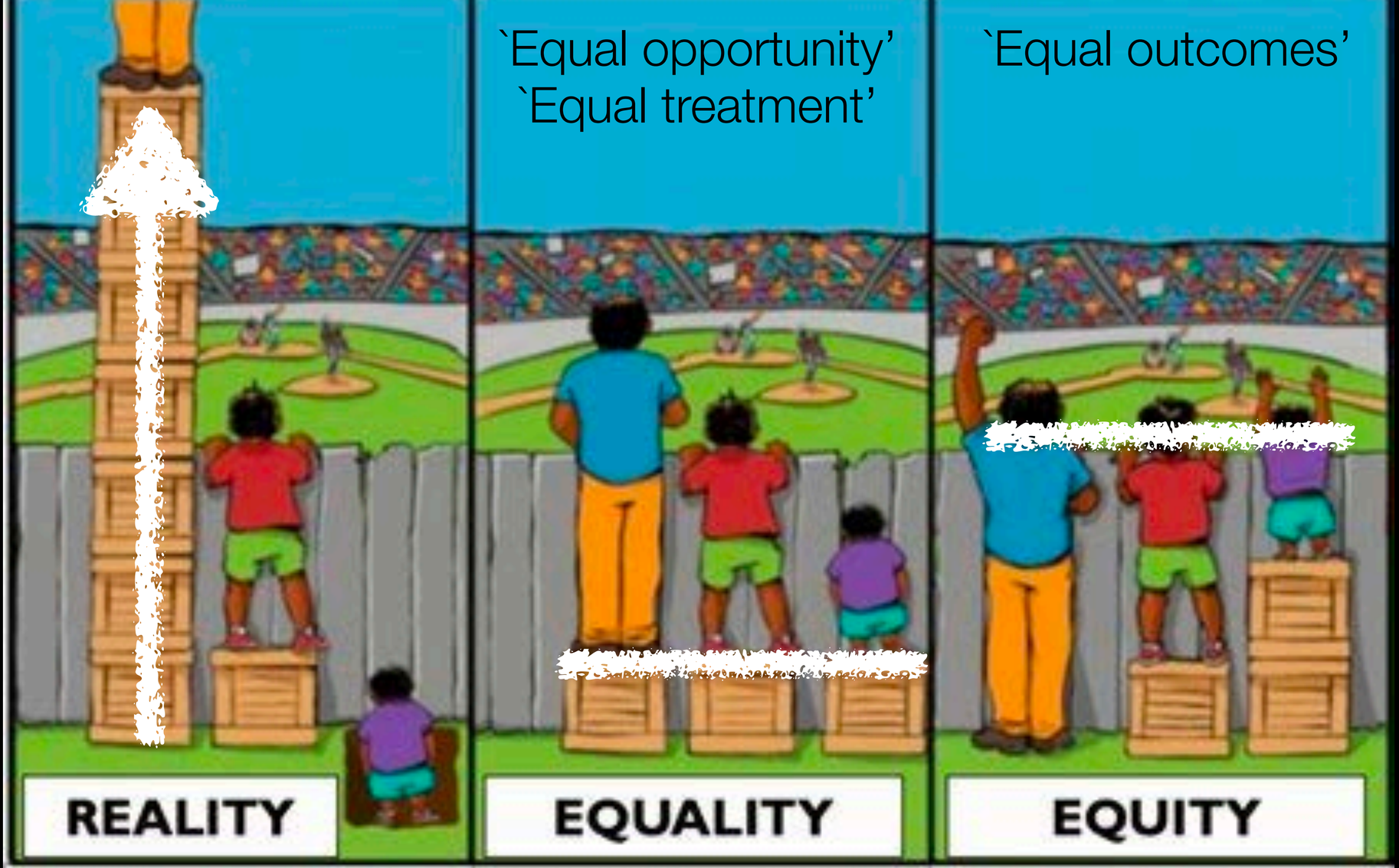

Annual Review of Statistics and Its Application

Algorithmic Fairness: Choices, Assumptions, and Definitions

Shira Mitchell,[1] Eric Potash,[2] Solon Barocas,[3,4] Alexander D'Amour,[5] and Kristian Lum[6]

`Equal opportunity’
`Equal treatment’

`Equal outcomes’



REALITY

One gets **more than** is needed, while the other gets **less than** is needed. Thus, a huge disparity is created.

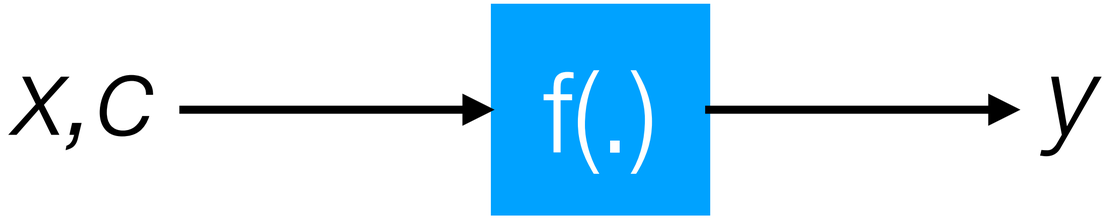
EQUALITY

The assumption is that **everyone benefits from the same supports**. This is considered to be equal treatment.

EQUITY

Everyone gets the support they need, which produces equity.

In equations



x = action, state

c = protected attribute

y = reward, outcome

$$p(y = 1|x, c) = p(y = 1|x)$$

$$p(y = 1|x, c = 0) = p(y = 1|x, c = 1)$$

Examples

New York City marathon

Men

2010	Gebregziabher Gebremariam	Ethiopia	2:08:14
2011	Geoffrey Mutai	Kenya	2:05:05
2012	**Cancelled** due to Hurricane Sandy		
2013	Geoffrey Mutai (2)	Kenya	2:08:24
2014	Wilson Kipsang	Kenya	2:10:59
2015	Stanley Biwott	Kenya	2:10:34
2016	Ghirmay Ghebreslassie	Eritrea	2:07:51
2017	Geoffrey Kamworor	Kenya	2:10:53
2018	Lelisa Desisa	Ethiopia	2:05:59
2019	Geoffrey Kamworor (2)	Kenya	2:08:13

Women

2010	Edna Kiplagat	Kenya	2:28:20
2011	Firehiwot Dado	Ethiopia	2:23:15
2012	**Cancelled** due to Hurricane Sandy		
2013	Priscah Jeptoo	Kenya	2:25:07
2014	Mary Keitany	Kenya	2:25:07
2015	Mary Keitany	Kenya	2:24:25
2016	Mary Keitany	Kenya	2:24:26
2017	Shalane Flanagan	United States	2:26:53
2018	Mary Keitany	Kenya	2:22:48
2019	Joyciline Jepkosgei	Kenya	2:22:38

Jocelyn Jepkosgei and Geoffrey Kamworor, winners of the 2019 NYC marathon

Dressage

Games	Gold	Silver	Bronze
1980 Moscow *details*	Elisabeth Theurer on *Mon Cheri* (AUT)	Yuri Kovshov on *Igrok* (URS)	Viktor Ugryumov on *Shalun* (URS)
1984 Los Angeles *details*	Reiner Klimke on *Ahlerich* (FRG)	Anne Grethe Jensen on *Marzog* (DEN)	Otto Hofer on *Limandus* (SUI)
1988 Seoul *details*	Nicole Uphoff on *Rembrandt* (FRG)	Margit Otto-Crépin on *Corlandus* (FRA)	Christine Stückelberger on *Chagall de Lully CH* (SUI)
1992 Barcelona *details*	Nicole Uphoff on *Rembrandt* (GER)	Isabell Werth on *Gigolo* (GER)	Klaus Balkenhol on *Goldstern* (GER)
1996 Atlanta *details*	Isabell Werth on *Gigolo* (GER)	Anky van Grunsven on *Bonfire* (NED)	Sven Rothenberger on *Weyden* (NED)
2000 Sydney *details*	Anky van Grunsven on *Bonfire* (NED)	Isabell Werth on *Gigolo* (GER)	Ulla Salzgeber on *Rusty* (GER)
2004 Athens *details*	Anky van Grunsven on *Salinero* (NED)	Ulla Salzgeber on *Rusty* (GER)	Beatriz Ferrer-Salat on *Beauvalais* (ESP)
2008 Beijing *details*	Anky van Grunsven on *Salinero* (NED)	Isabell Werth on *Satchmo* (GER)	Heike Kemmer on *Bonaparte* (GER)
2012 London *details*	Charlotte Dujardin on *Valegro* (GBR)	Adelinde Cornelissen on *Parzival* (NED)	Laura Bechtolsheimer on *Mistral Højris (GBR)*
2016 Rio *details*	Charlotte Dujardin on *Valegro* (GBR)	Isabell Werth on *Weihegold* (GER)	Kristina Bröring-Sprehe on *Desperados* (GER)

Jumping

Games	Gold	Silver	Bronze
1980 Moscow *details*	Jan Kowalczyk on *Artemor* (POL)	Nikolai Korolkov on *Espadron* (URS)	Joaquín Perez Heras on *Alymony* (MEX)
1984 Los Angeles *details*	Joseph Fargis on *Touch of Class* (USA)	Conrad Homfeld on *Abdullah* (USA)	Heidi Robbiani on *Jessica V* (SUI)
1988 Seoul *details*	Pierre Durand Jr. on *Jappeloup* (FRA)	Greg Best on *Gem Twist* (USA)	[illegible] on *Nepomuk* (FRG)
1992 Barcelona *details*	Ludger Beerbaum on *Classic Touch* (GER)	Piet Raymakers on *Ratina Z* (NED)	Norman Dello Joio on *Irish* (USA)
1996 Atlanta *details*	Ulrich Kirchhoff on *Jus de Pommes* (GER)	Willi Melliger on *Calvaro V* (SUI)	Alexandra Ledermann on *Rochet M* (FRA)
2000 Sydney *details*	Jeroen Dubbeldam on *De Sjiem* (NED)	Albert Voorn on *Lando* (NED)	Khaled Al-Eid on *Khashm Al Aan* (KSA)
2004 Athens *details*	Rodrigo Pessoa on *Baloubet du Rouet* (BRA)	Chris Kappler on *Royal Kaliber* (USA)	Marco Kutscher on *Montender* (GER)
2008 Beijing *details*	Eric Lamaze on *Hickstead* (CAN)	Rolf-Göran Bengtsson on *Ninja* (SWE)	Beezie Madden on *Authentic* (USA)
2012 London *details*	Steve Guerdat on *Nino Des Buissonets* (SUI)	Gerco Schroder on *London* (NED)	Cian O'Connor on *Blue Loyd* (IRL)
2016 Rio *details*	Nick Skelton on *Big Star* (GBR)	Peder Fredricson on *All In* (SWE)	Eric Lamaze on *Fine Lady* (CAN)

Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

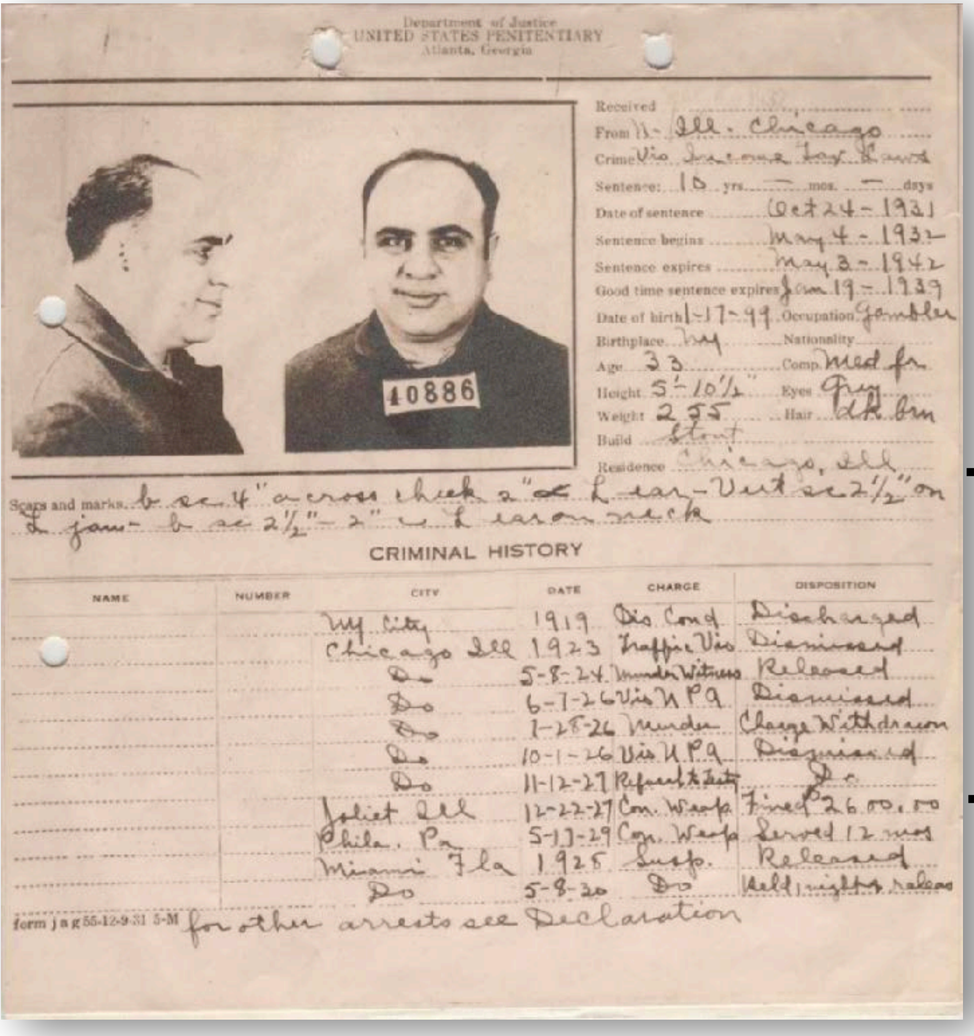
Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

Parole


Department of Justice
UNITED STATES PENITENTIARY
Atlanta, Georgia

40886

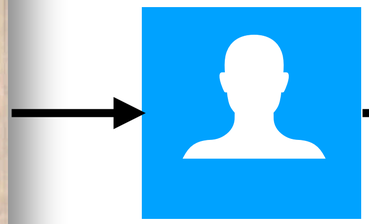
Received
From Ill. Chicago
Crime Vio Income Tax Laws
Sentence: 10 yrs. — mos. — days
Date of sentence Oct 24 - 1931
Sentence begins May 4 - 1932
Sentence expires May 3 - 1942
Good time sentence expires Jan 19 - 1939
Date of birth 1-17-99 Occupation Gambler
Birthplace Ny Nationality
Age 33 Comp. Med fr
Height 5'-10½" Eyes Gray
Weight 255 Hair Dk Brn
Build Stout
Residence Chicago, Ill

Scars and marks b ear 4" across chest 2" ot L ear-Vert ac 2½" on L jaw- b ac 2½" - 2" v t ear on neck

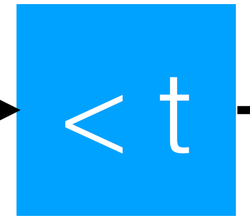
CRIMINAL HISTORY

NAME	NUMBER	CITY	DATE	CHARGE	DISPOSITION
		My City	1919	Dis Cond	Discharged
		Chicago Ill	1923	Traffic Vio	Dismissed
		Do	5-8-24	Murder Witness	Released
		Do	6-1-26	Vio N P A	Dismissed
		Do	1-28-26	Murder	Charge Withdrawn
		Do	10-1-26	Vio N P A	Dismissed
		Do	11-12-27	Refused State	Do
		Joliet Ill	12-22-27	Con. Weap	Fined $26.00 & Cost
		Phila. Pa	5-13-29	Con. Weap	Served 12 mos
		Miami Fla	1925	Susp.	Released
		Do	5-8-30	Do	Held 1 night & released

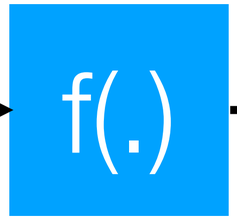
form j a g 55-12-9 M 5-M for other arrests see Declaration



Probability of reoffending



Free

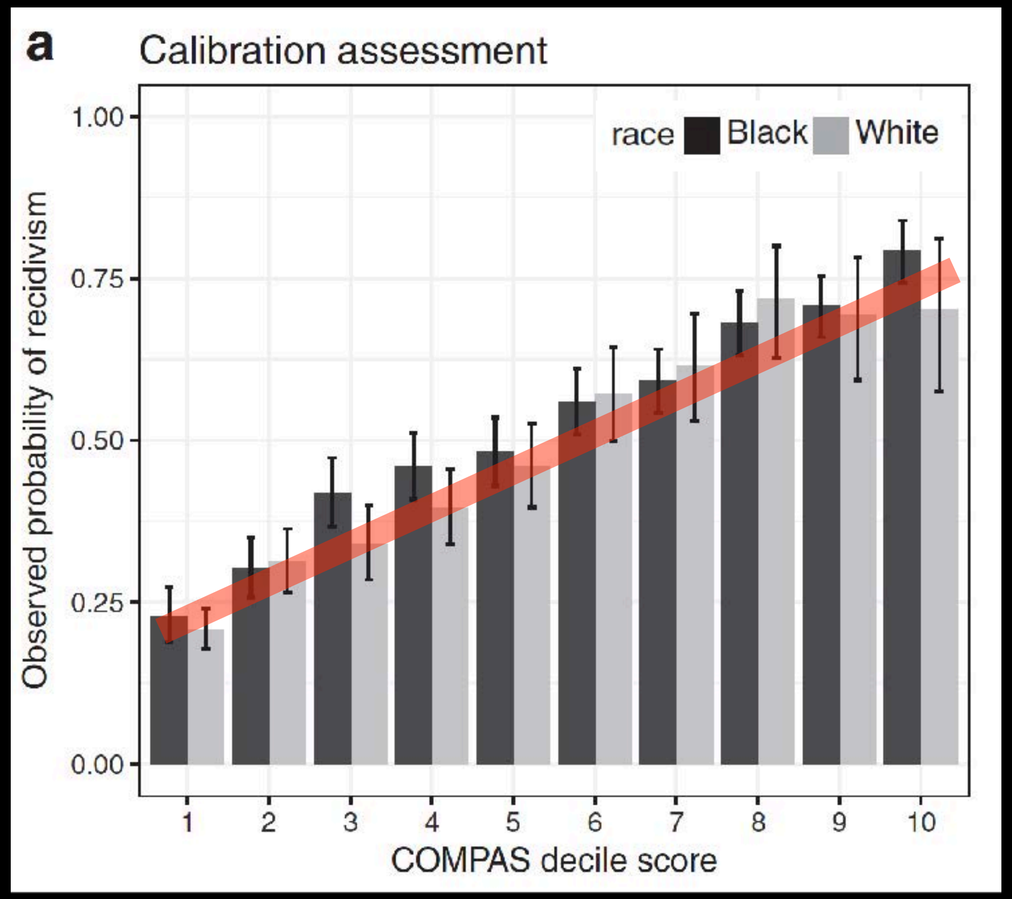


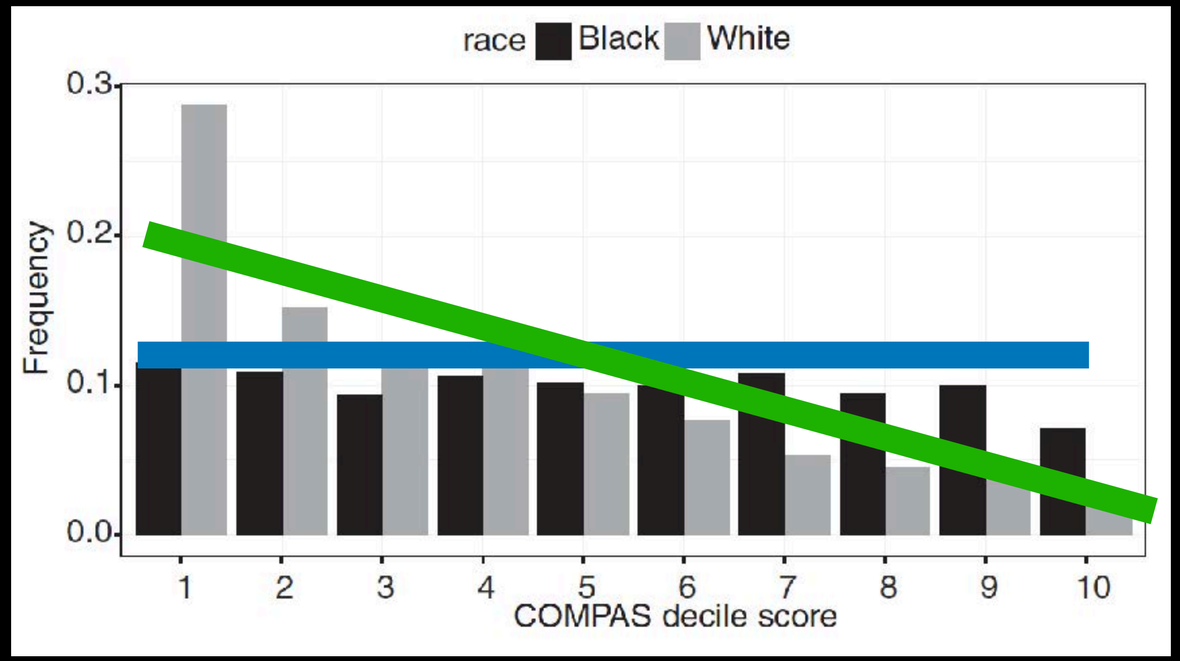
Risk 1-10

Reoffend? Y/N

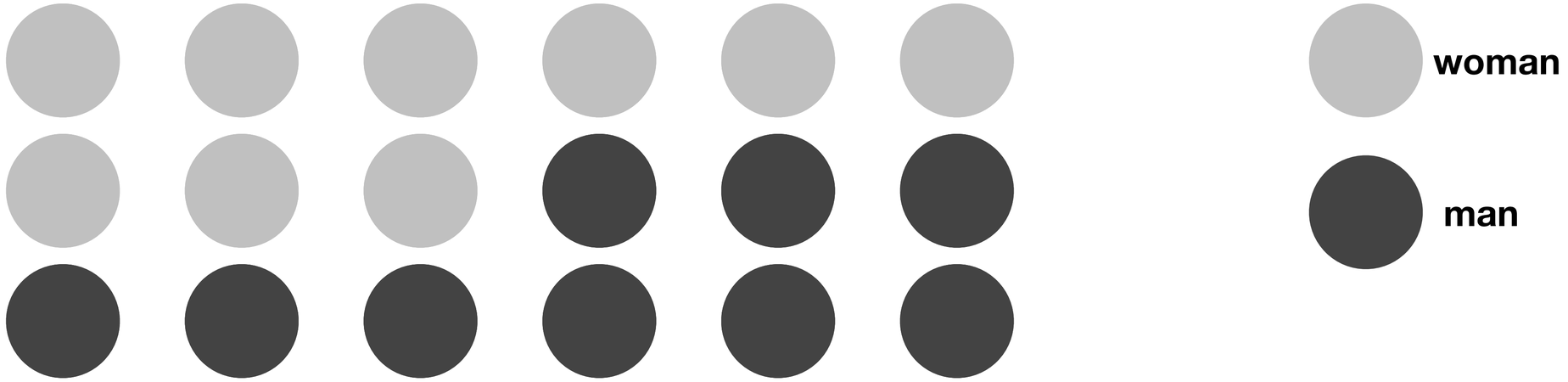
Prediction Fails Differently for Black Defendants [Pro-Publica 2015]

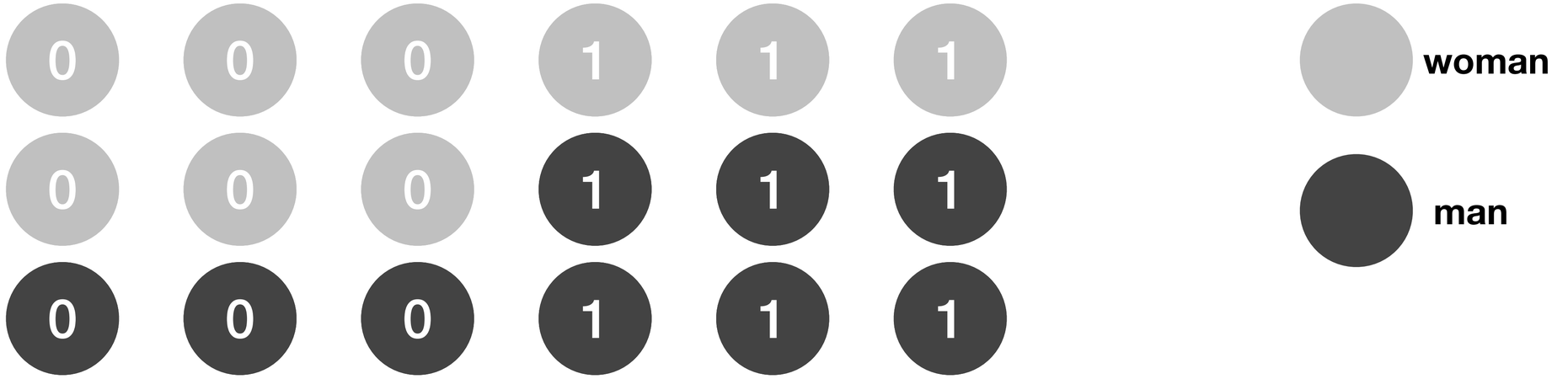
	WHITE	AFRICAN AMERICAN
Non-reoffenders - % labeled hi risk	23.5%	44.9%
Reoffenders - % labeled lower risk	47.7%	28.0%

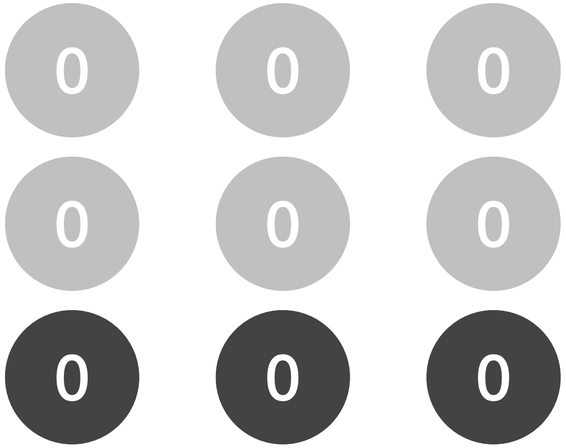




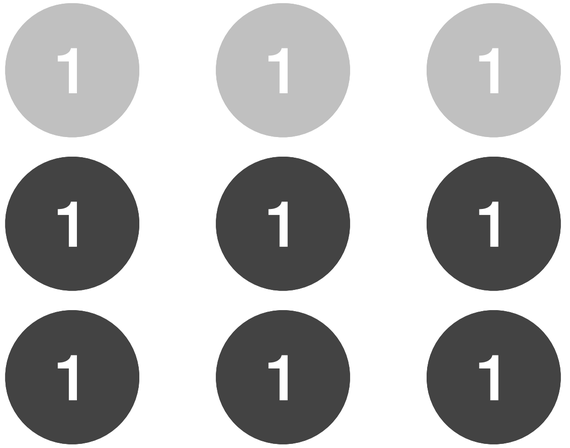
[Chouldechova 2017]


woman
man

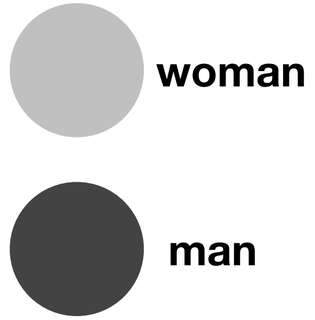

0 0 0 1 1 1
0 0 0 1 1 1
0 0 0 1 1 1
woman
man

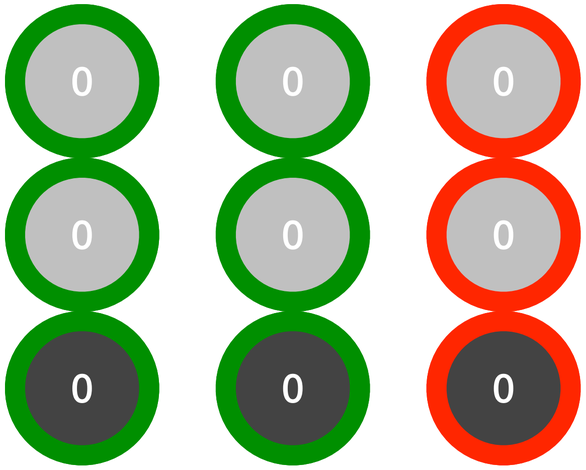


LOW SCORE

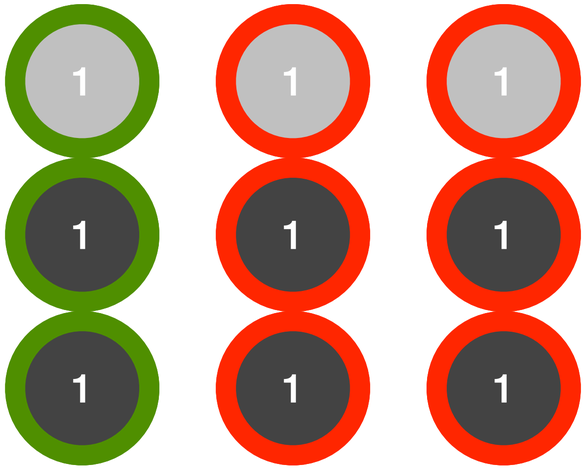


HIGH SCORE

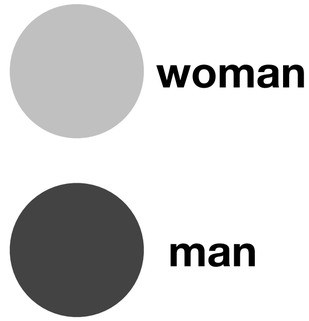




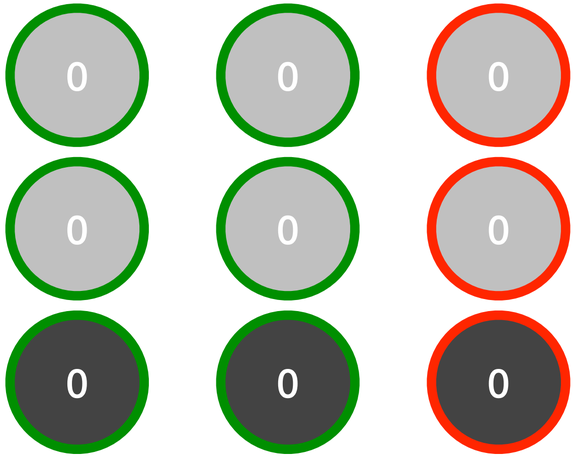
LOW SCORE



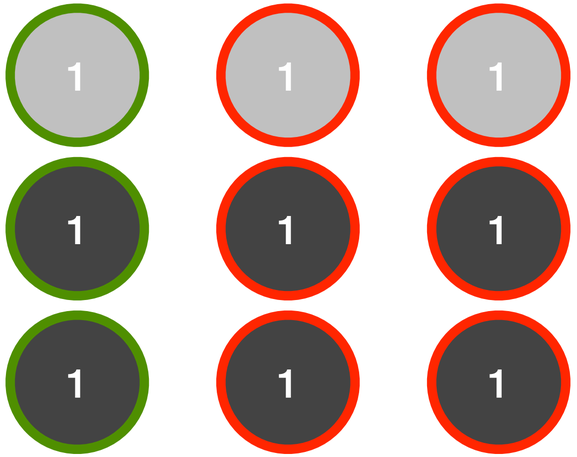
HIGH SCORE



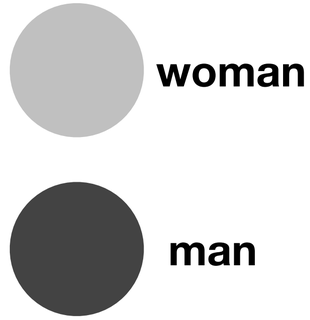
PREDICTION



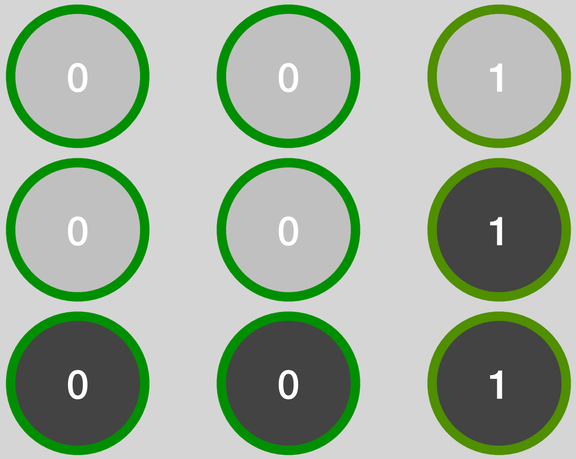
LOW SCORE



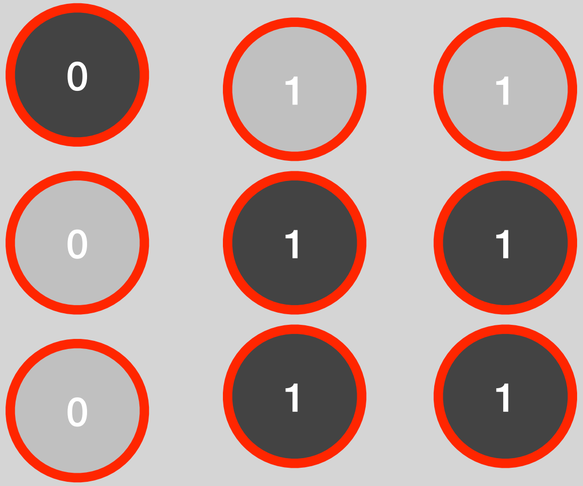
HIGH SCORE



POST-DICTION

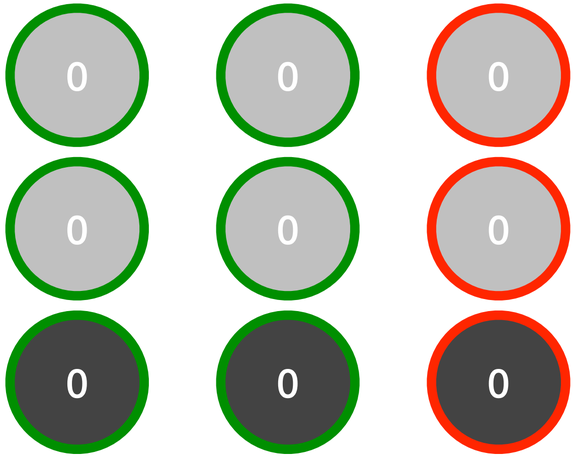


NO REOFFEND

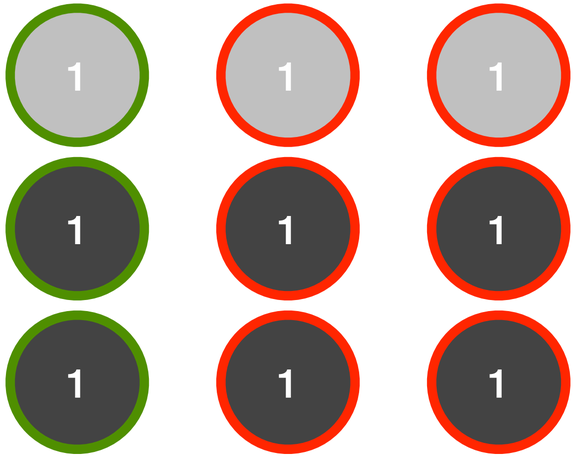


REOFFEND

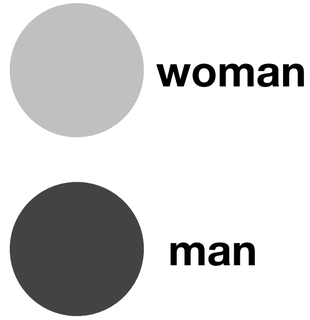
PREDICTION



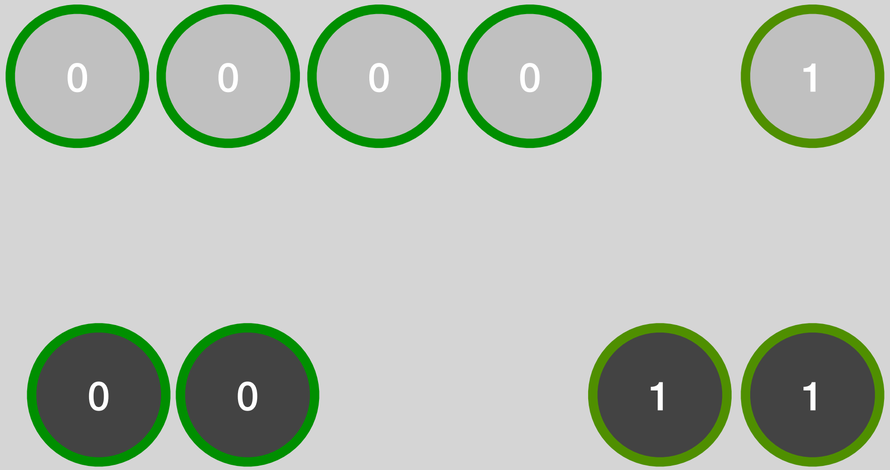
LOW SCORE



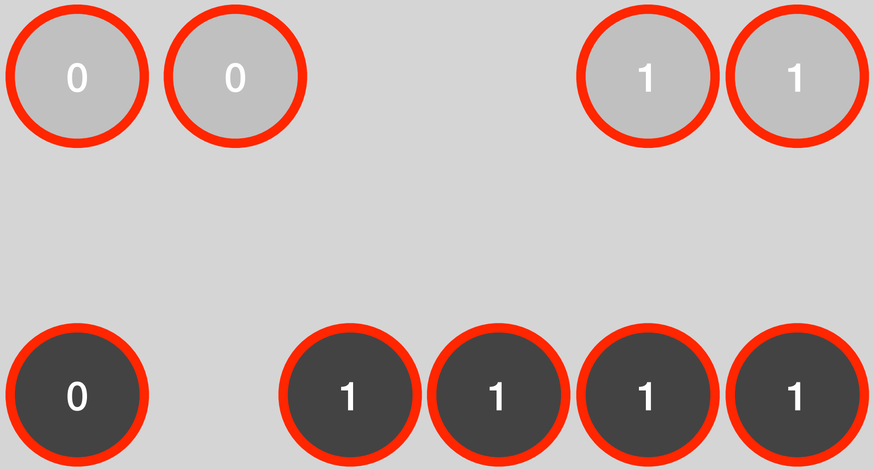
HIGH SCORE



POST-DICTION

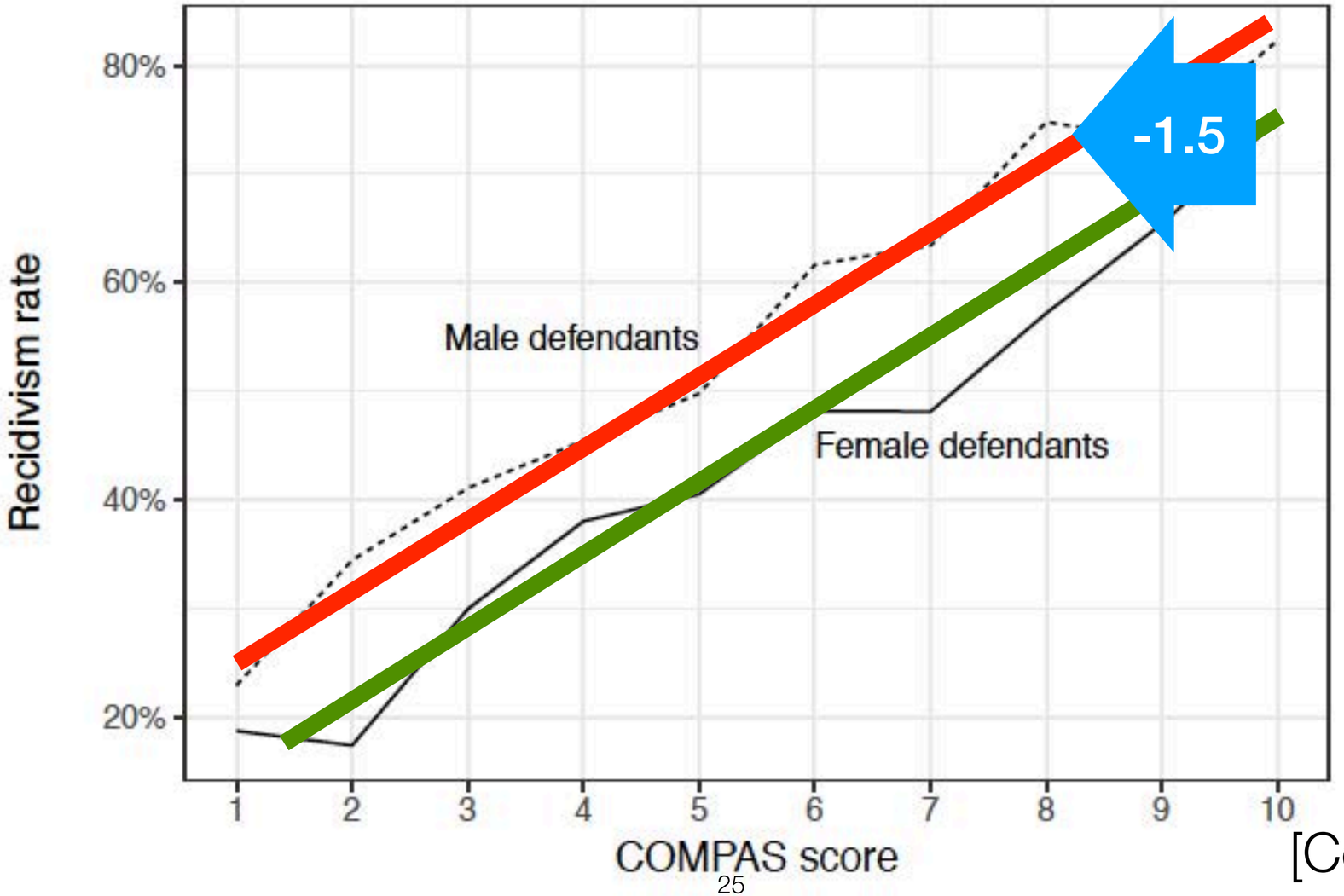


NO REOFFEND



REOFFEND

Male vs females…



[Corbett-Davies 2018]

Learnings

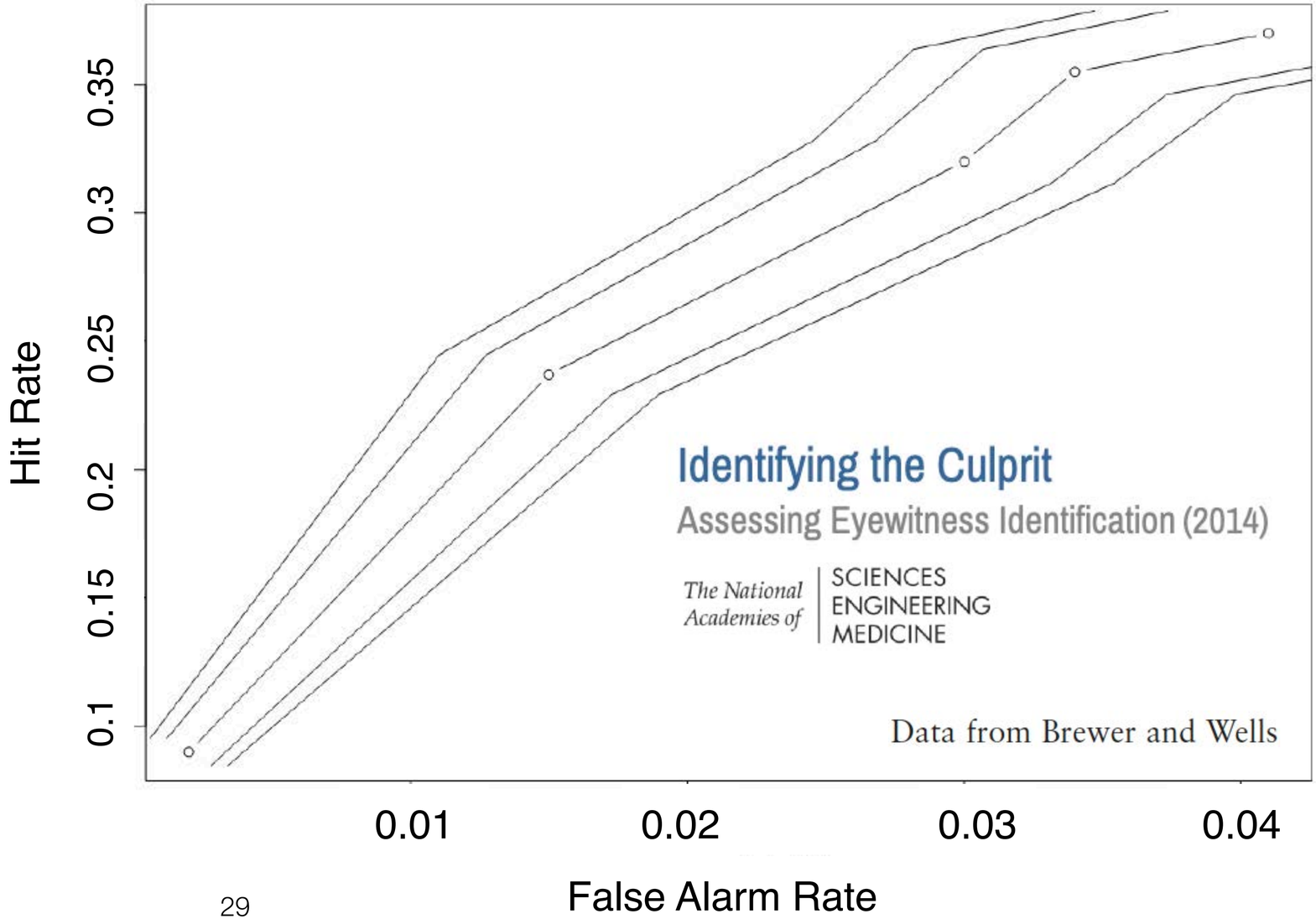
- Individual vs group
- Protected attributes
- Definitions of fairness
- Cannot be fair according to multiple definitions
- Society decides criterion, AI implements it

Face analysis
Face recognition


The New York Times
We Teach A.I. Systems Everything, Including Our Biases
The New York Times
Facial Recognition Is Accurate, if You're a White Guy
By Steve Lohr
Feb. 9, 2018
FAST COMPAN
TECH
WORK LIFE
CREATIVITY
IMPACT
AUDIO
COMPASS
Sign up for our daily email.
Enter your ema
05.22.19
Here's AOC calling out the vicious circle of white men building biased face AI

Human error

Ronald Cotton and Jennifer Thompson



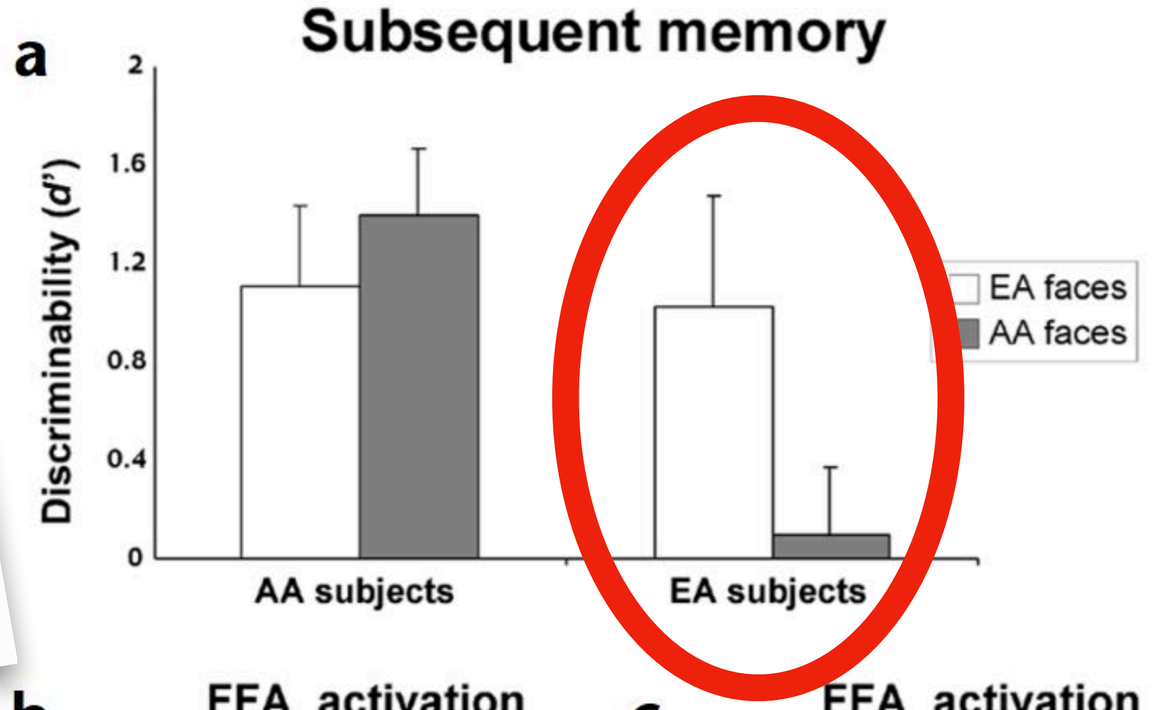
Human bias

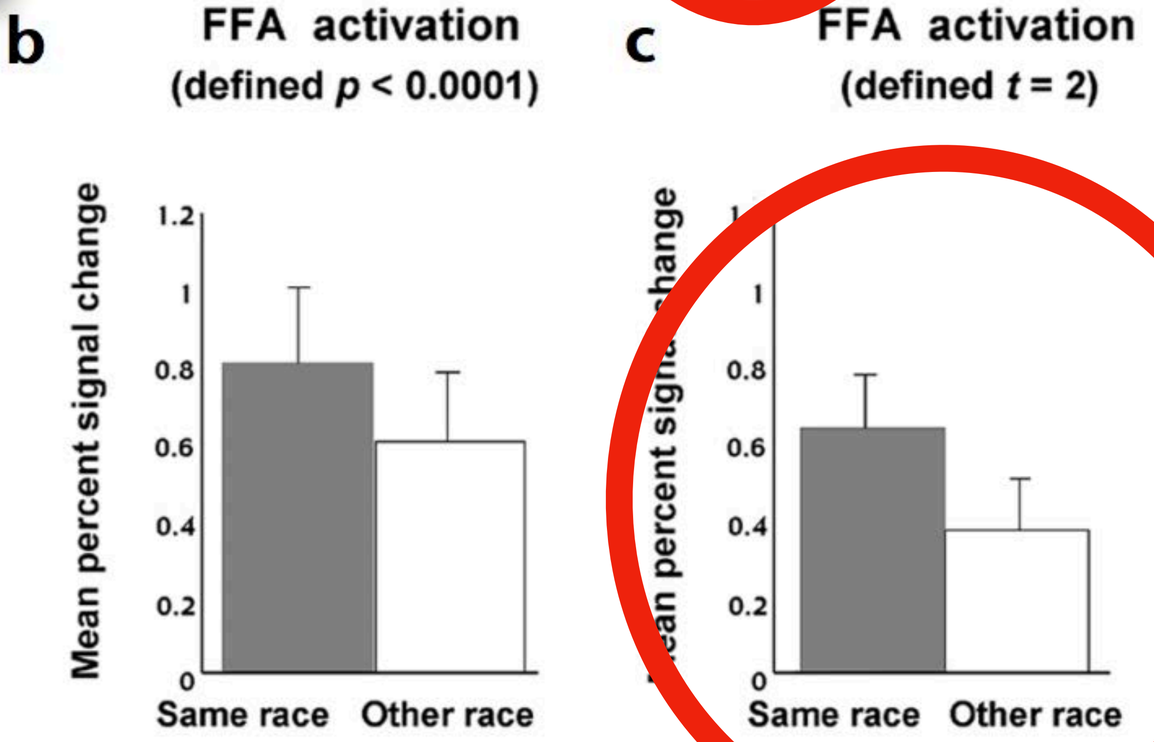
Differential responses in the fusiform region to same-race and other-race faces

Alexandra J. Golby, John D. E. Gabrieli, Joan Y. Chiao and Jennifer L. Eberhardt

Departments of Radiology and Psychology, Jordan Hall-Building 420, Stanford University, Stanford, California 94305, USA
Correspondence should be addressed to J.L.E. (jle@psych.stanford.edu)

nature neuroscience • volume 4 no 8 • august 2001





The New York Times

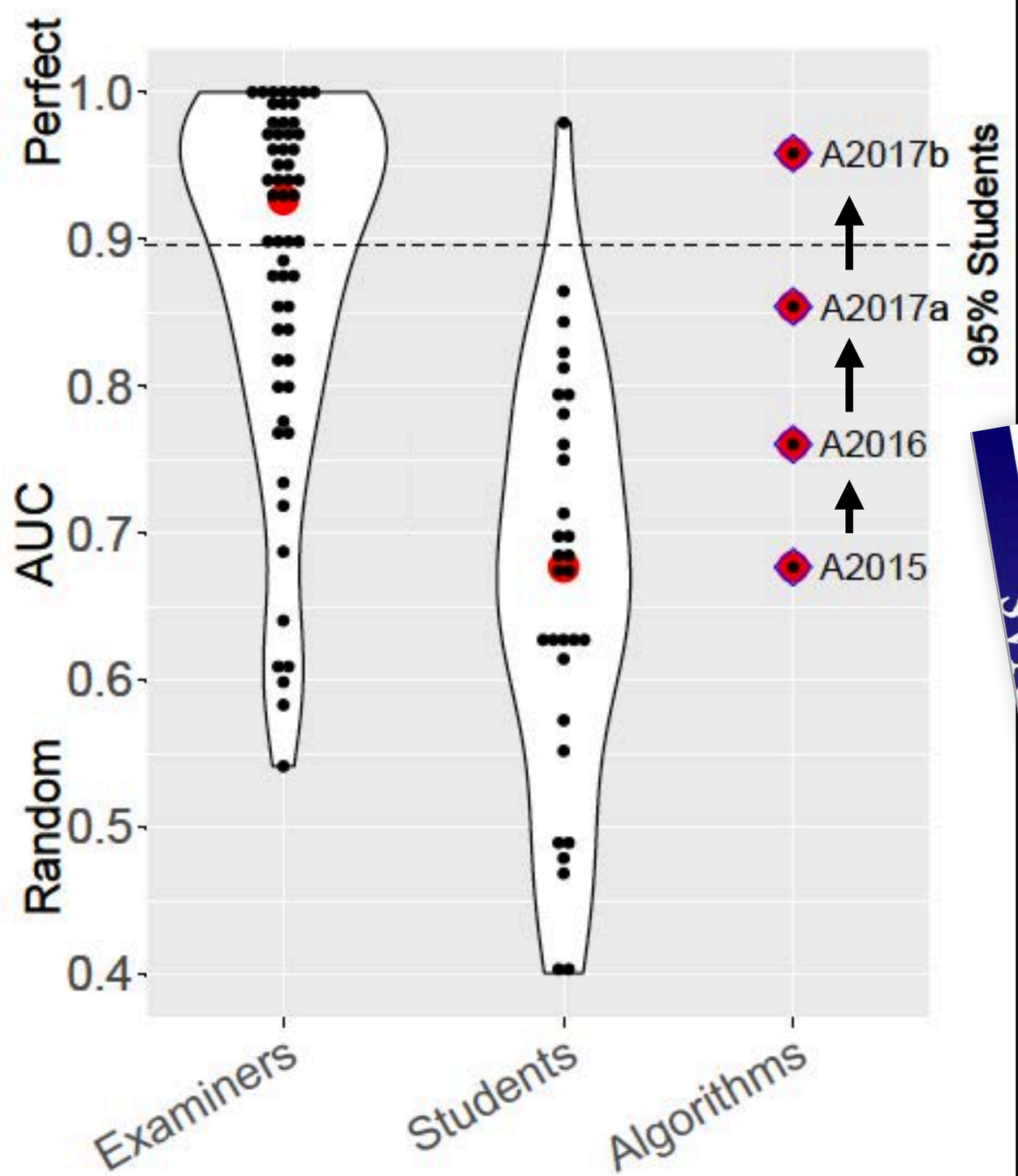
ECONOMIC VIEW

Biased Algorithms Are Easier to Fix Than Biased People

Racial discrimination by algorithms or by people is harmful — but that's where the similarities end.

By Sendhil Mullainathan

Dec. 6, 2019



Accuracy, consistency

PNAS

Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms

P. Jonathon Phillips[a,1], Amy N. Yates[a], Ying Hu[b], Carina A. Hahn[b], Eilidh Noyes[b], Kelsey Jackson[b], Jacqueline G. Cavazos[b], Géraldine Jeckeln[b], Rajeev Ranjan[c], Swami Sankaranarayanan[c], Jun-Cheng Chen[d], Carlos D. Castillo[d], Rama Chellappa[c], David White[e], and Alice J. O'Toole[b]

[a]Information Access Division, National Institute of Standards and Technology, Gaithersburg, MD 20899; [b]School of Behavioral and Brain Sciences, The University of Texas at Dallas, Richardson, TX 75080; [c]Department of Electrical and Computer Engineering, University of Maryland Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20854; [d]University of Maryland Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20854; and [e]School of Psychology, The University of New South Wales, Sydney, NSW 2052, Australia


TES OF AMERICA
U.S. CUSTOMS AND BORDER PROTECTION

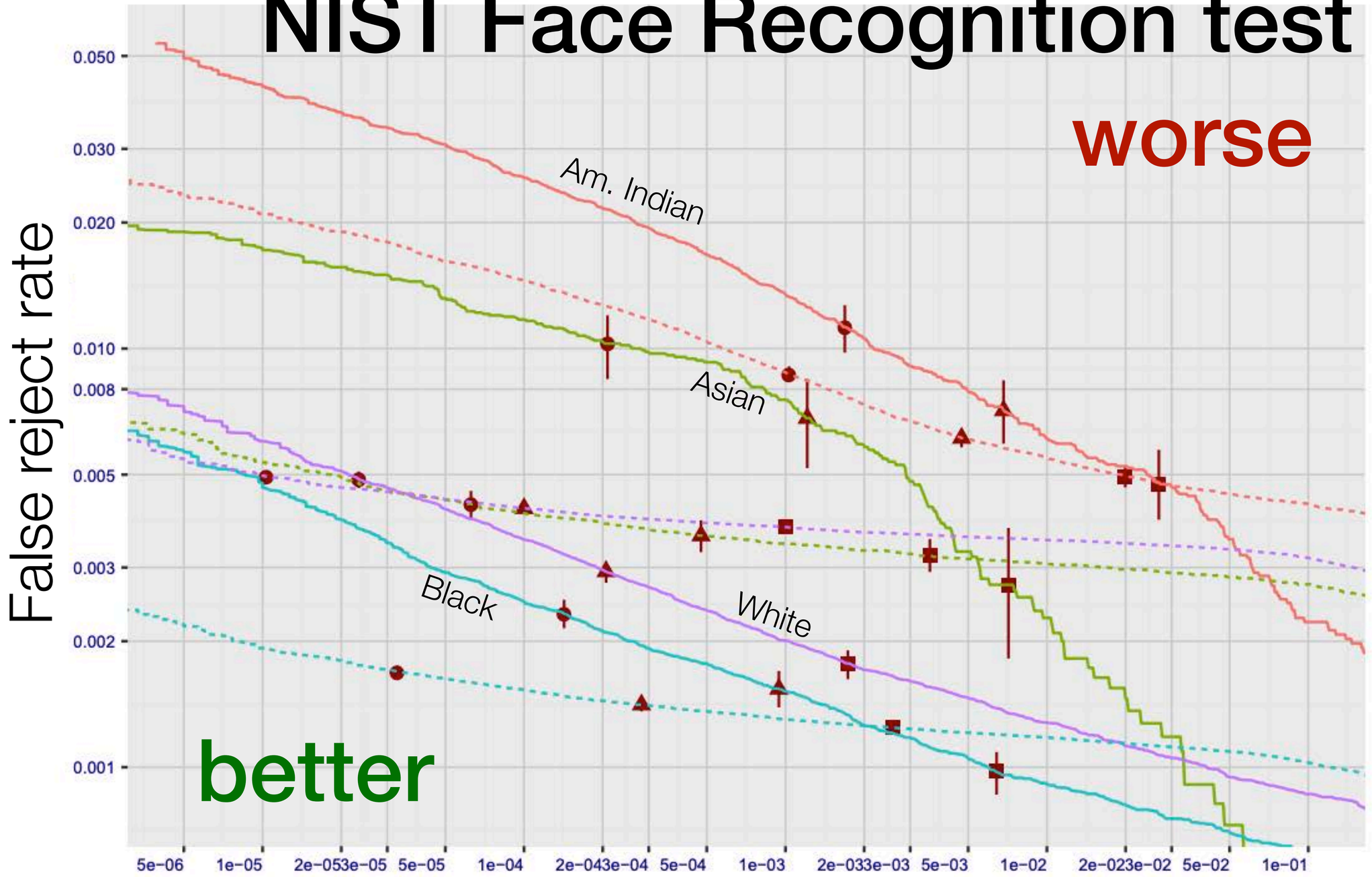

GLOBAL ENTRY
U.S. Customs and Border Protection

Benefits

- Cost
- Speed
- Scalability
- Fairness
- Accuracy
- Accountability
- Improvability
- Consistency

NIST Face Recognition test

(MUGSHOT)



[NIST 2019]

Race Bias ?

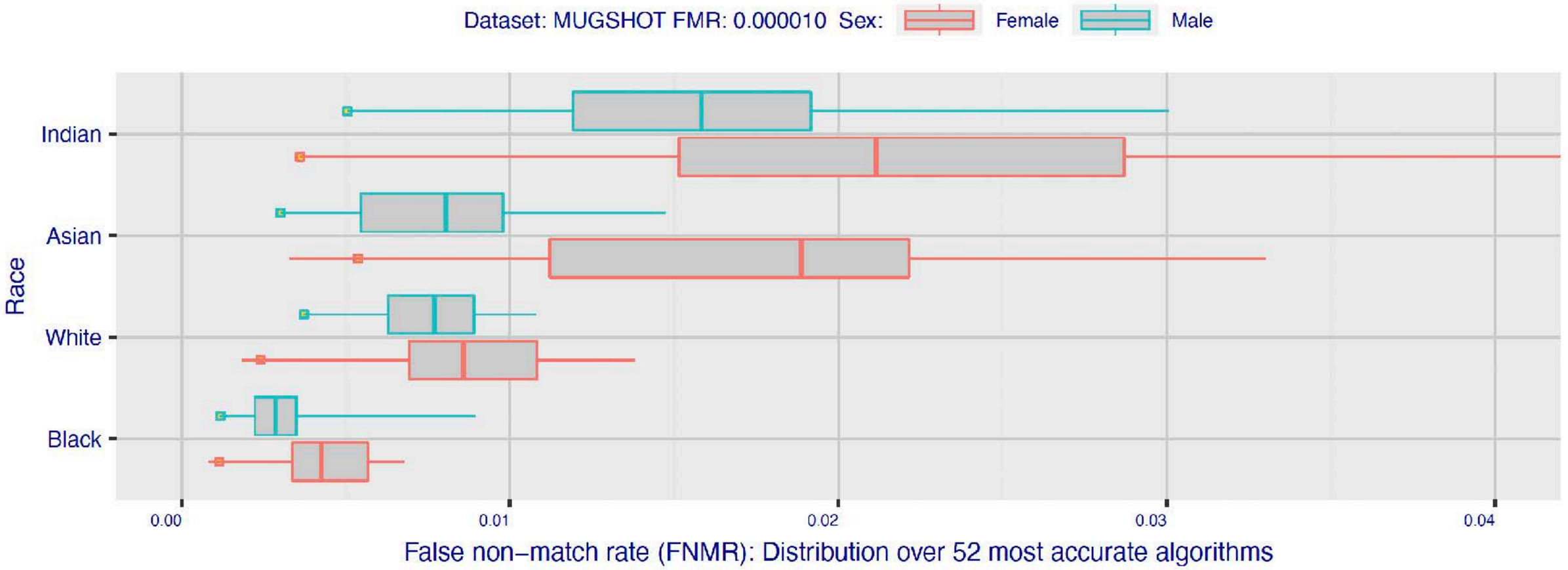
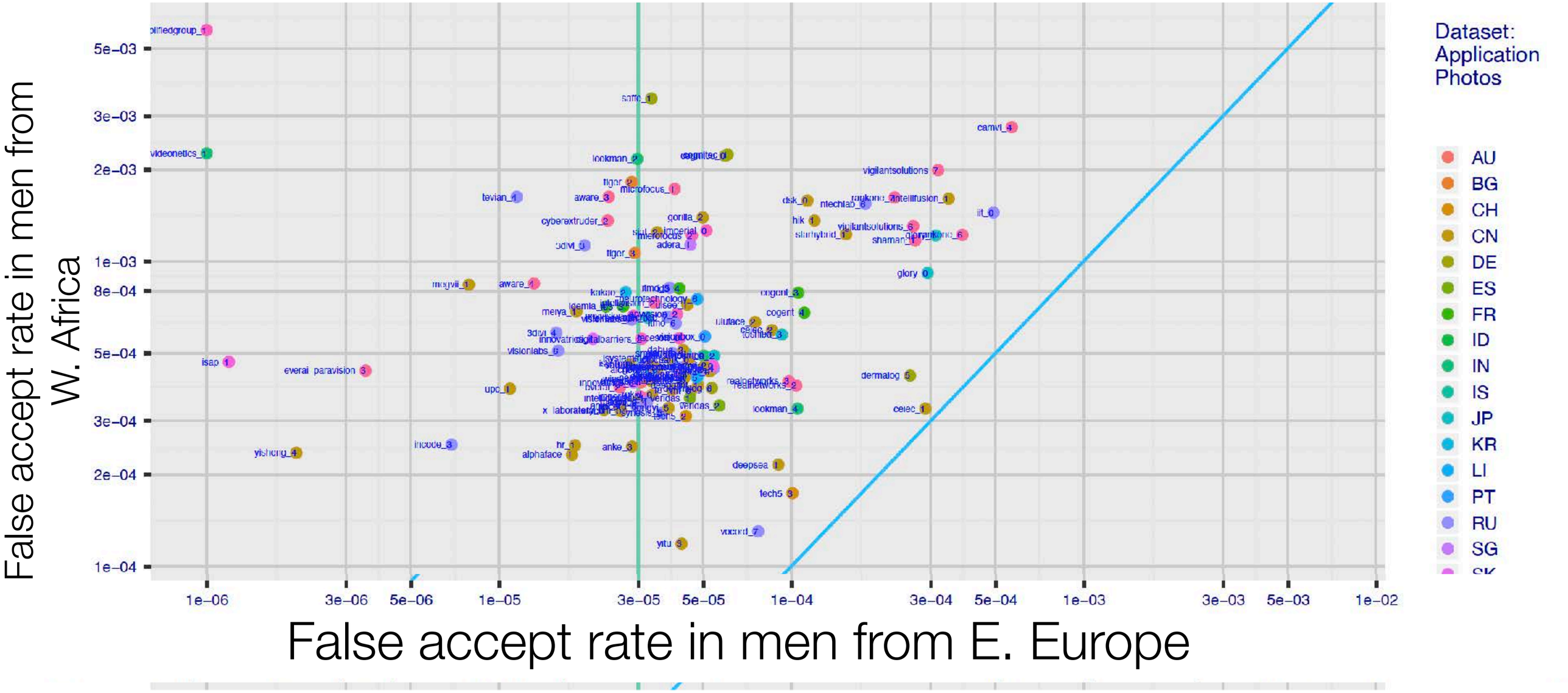
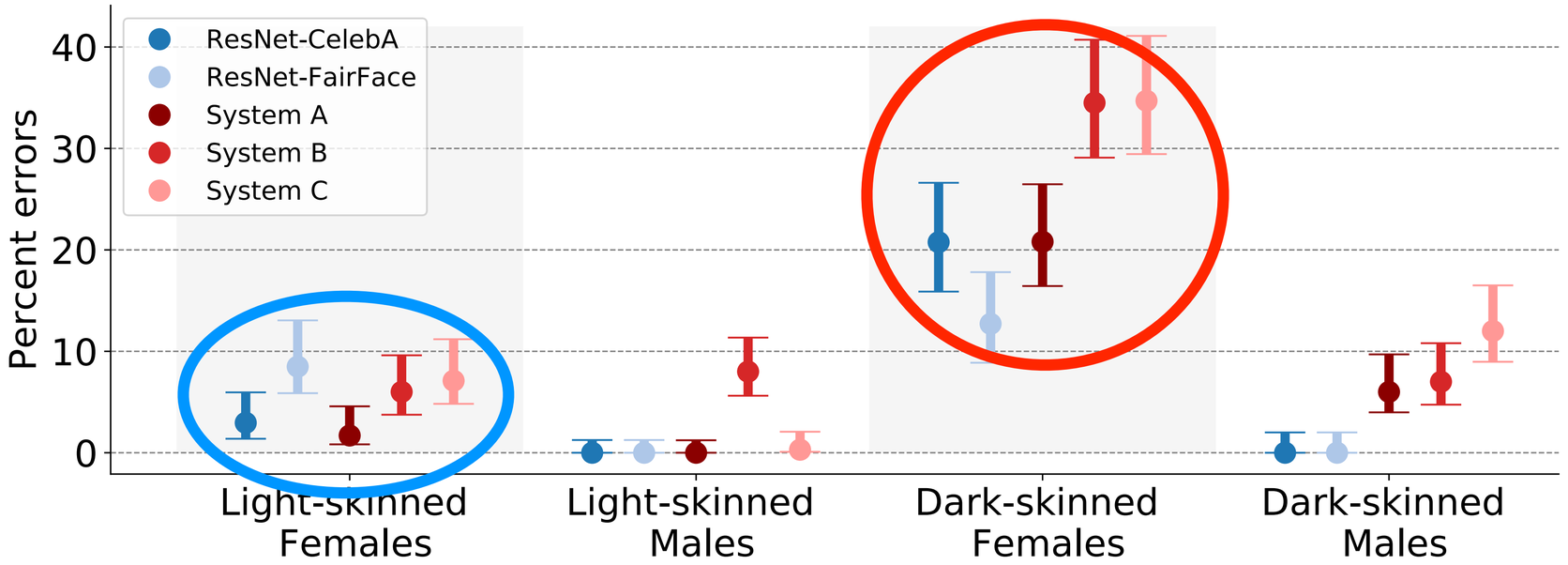


Figure 17: For mugshot comparisons, the figure shows the distribution of FNMR values over the 52 most accurate verification algorithms, by sex and race. The threshold was set for each algorithm to achieve FMR = 0.00001 over all imposter comparisons. The line within each box is the median over those algorithms; the box itself spans the interquartile range (26 algorithms) and the lines here extend to minimum and maximum values. The small box on the left side indicates the accuracy for best algorithm overall, on this dataset alphaface001.

Race bias ?

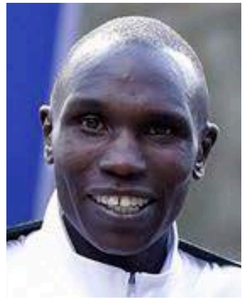

False accept rate in men from W. Africa
False accept rate in men from E. Europe
5e−03
3e−03
2e−03
1e−03
8e−04
5e−04
3e−04
2e−04
1e−04
1e−06
3e−06
5e−06
1e−05
3e−05
5e−05
1e−04
3e−04
5e−04
1e−03
3e−03
5e−03
1e−02
Dataset: Application Photos
AU
BG
CH
CN
DE
ES
FR
ID
IN
IS
JP
KR
LI
PT
RU
SG
camvi_4
videonetics_1
vigilantsolutions_7
tiger_2
microfocus_1
tevian_1
aware_3
dsk_0
ntechlab_6
iit_0
cyberextruder_2
gorilla_2
hik_1
vigilantsolutions_6
imperial_0
adera_1
tiger_3
glory_0
megvii_1
aware_1
cogent_3
cogent_4
3divi_4
dermalog_5
isap_1
everai_paravision_3
realnetworks_3
realnetworks_2
upc_1
lookman_4
ceiec_1
yitu_5
incode_3
hr_1
anke_3
alphaface_1
yishong_4
deepsea_1
tech5_3
vocord_7

Gender classification



PPB dataset
(Scandinavia + Sub-Saharan Africa)
[Buolamwini & Gebru 2018]

Parliament web sites of Sweden, Finland and Iceland



Parliament web sites of Rwanda, Senegal and South Africa

Observational study

sample

annotate

algorithm

Female
Dark Skin

prediction

	Light skin	Dark skin
Female	3%	**25%**
Male	5%	3%

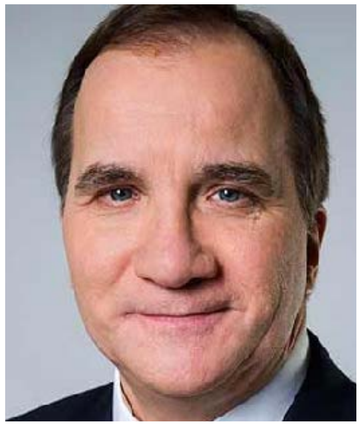
The New York Times

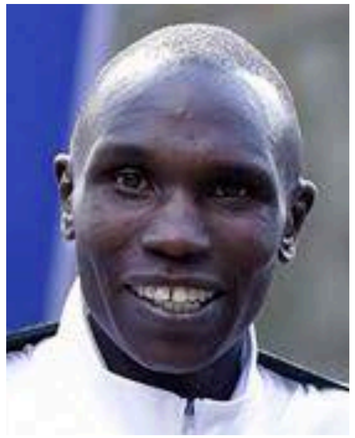
Facial Recognition Is Accurate, if You're a White Guy

By Steve Lohr

Feb. 9, 2018

COVID mortality

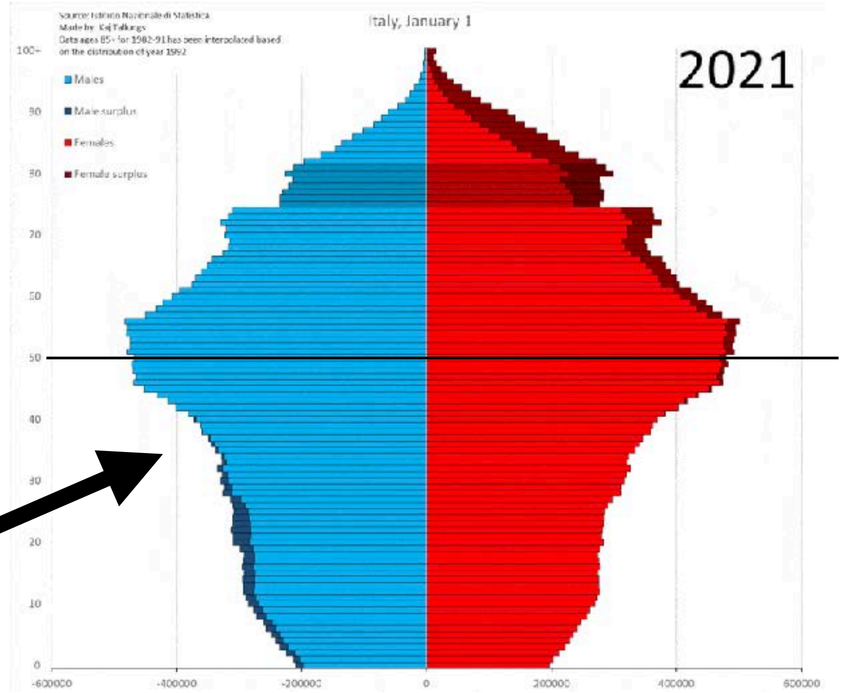


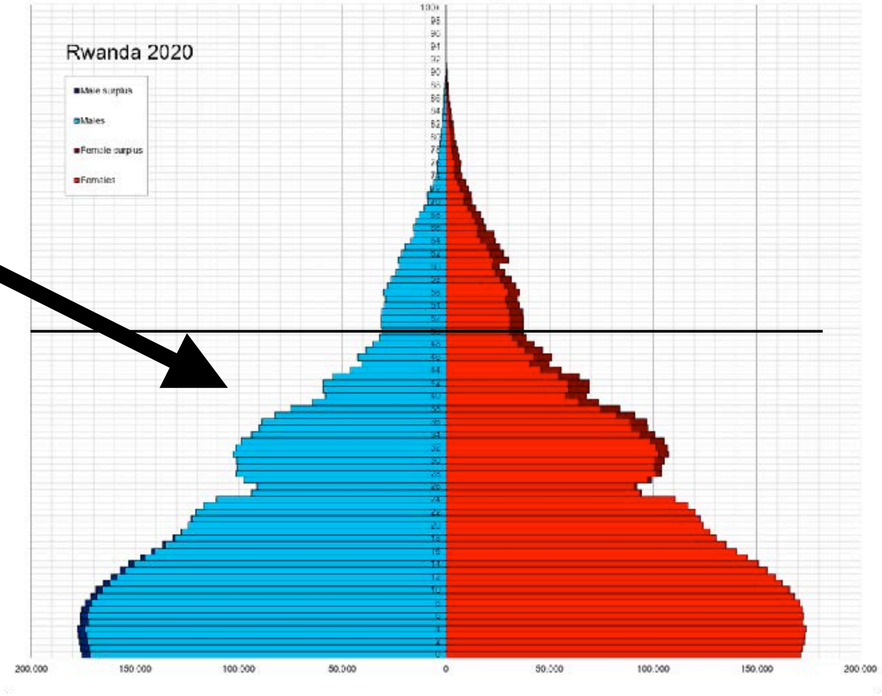


COVID MORTALITY

Country	Deaths/100K pop.
United Kingdom	
Poland	
Italy	
Sudan	
Rwanda	
Kenya	

[Johns Hopkins Coronavirus Resource Center, 2023]





Huge bias vis-a-vis ethnicity?

The NEW ENGLAND JOURNAL of MEDICINE

Chocolate Consumption, Cognitive Function, and Nobel Laureates

Franz H. Messerli, M.D.

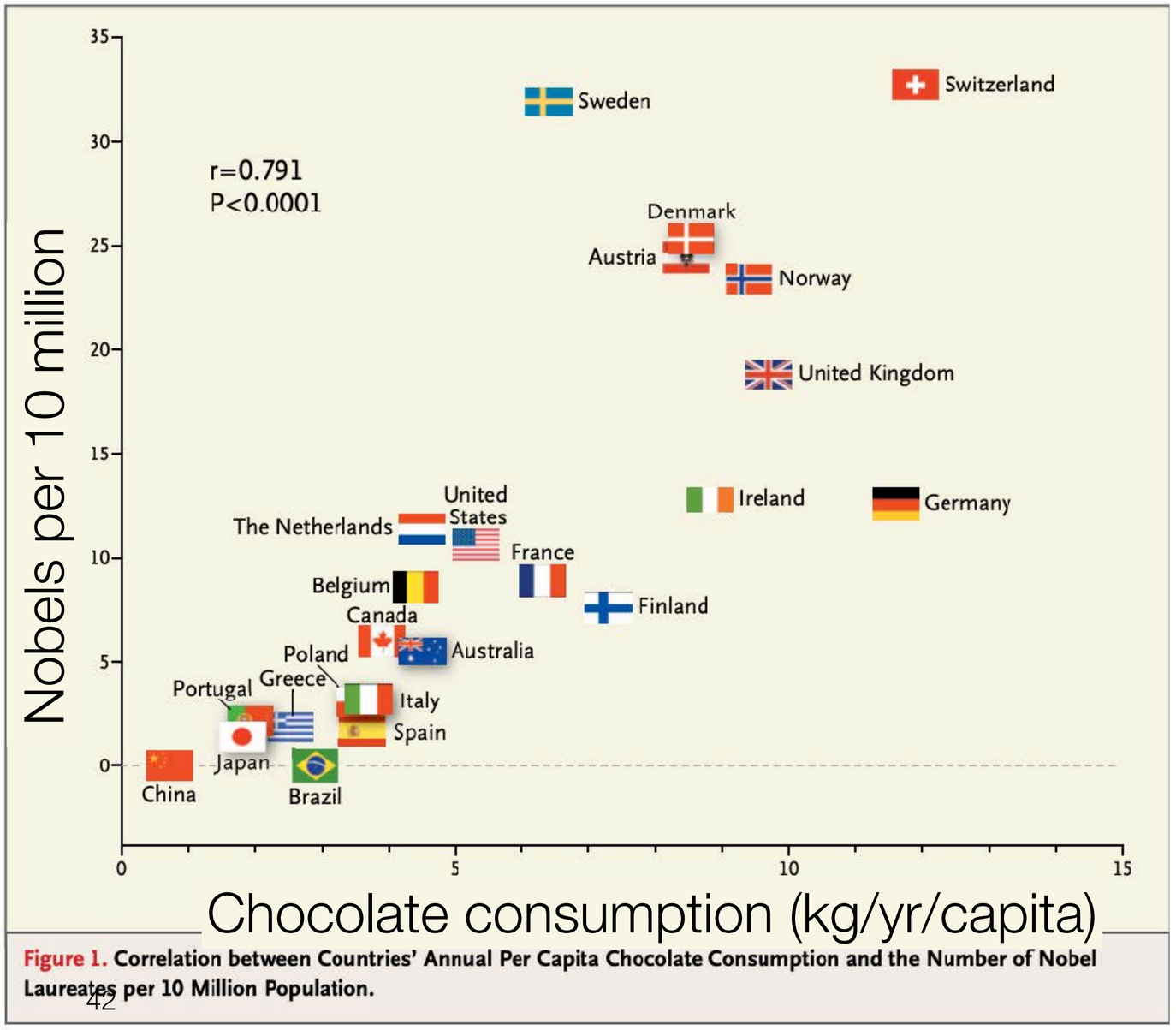
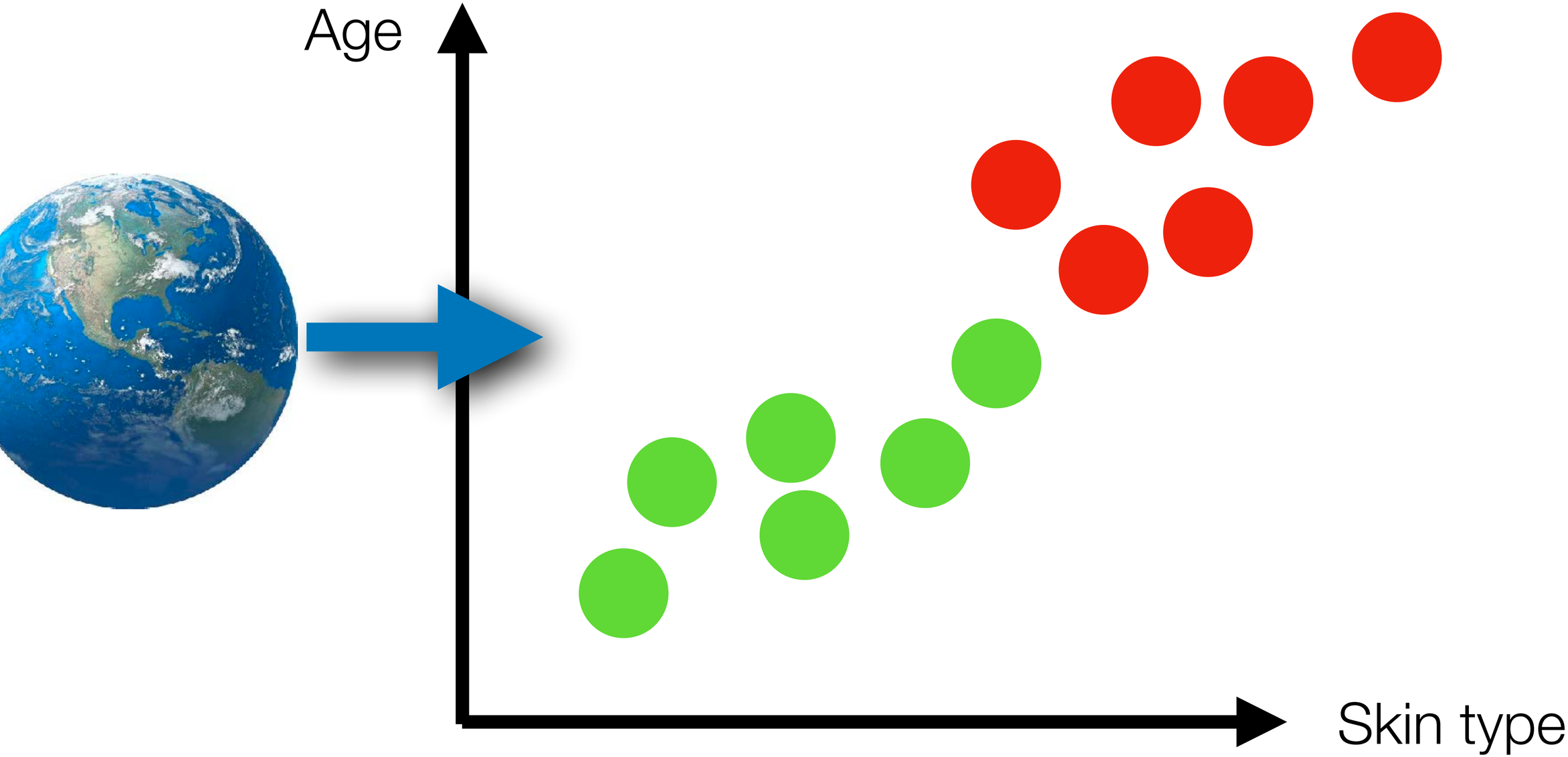
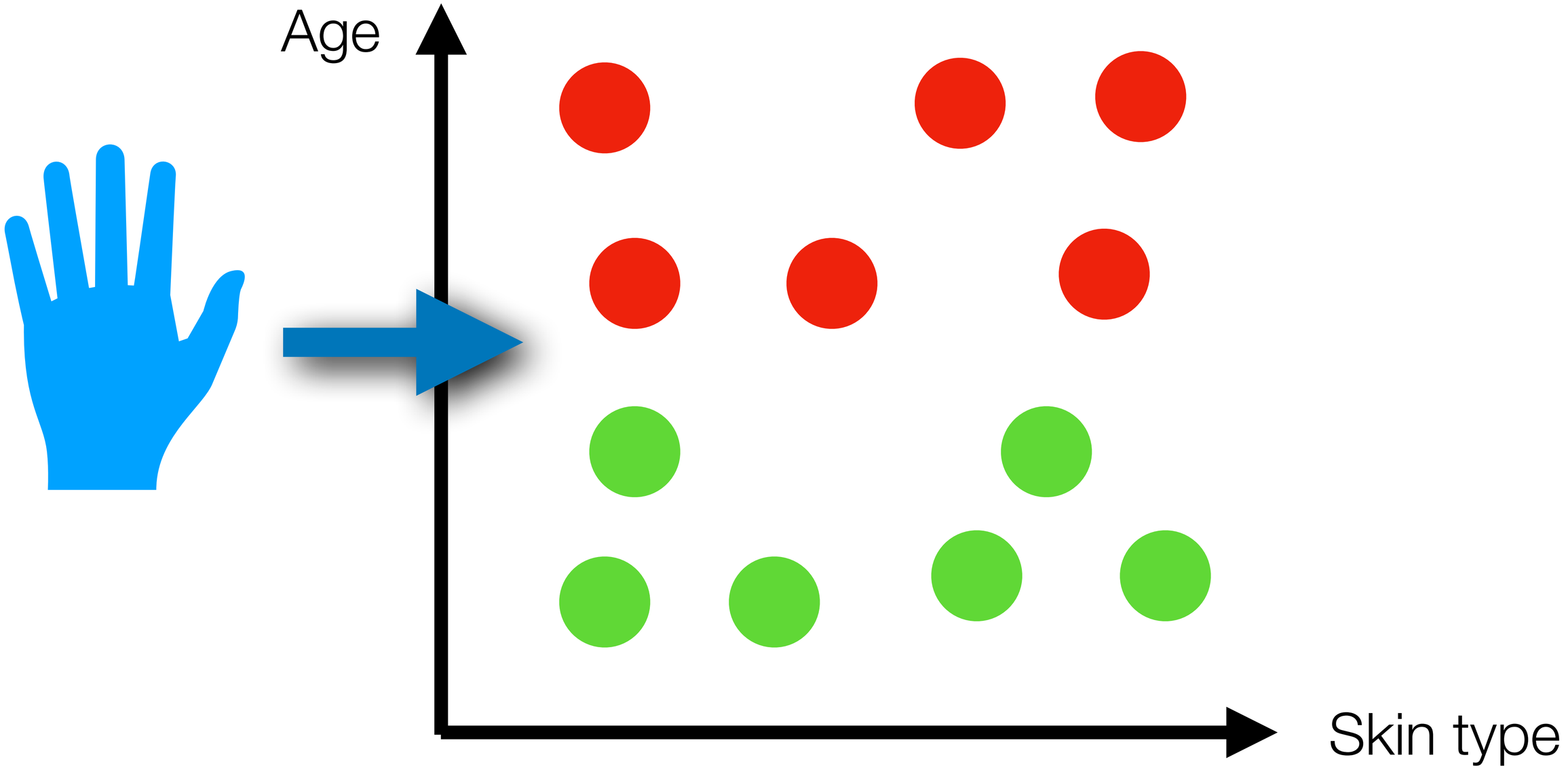


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Observational study



Experiment

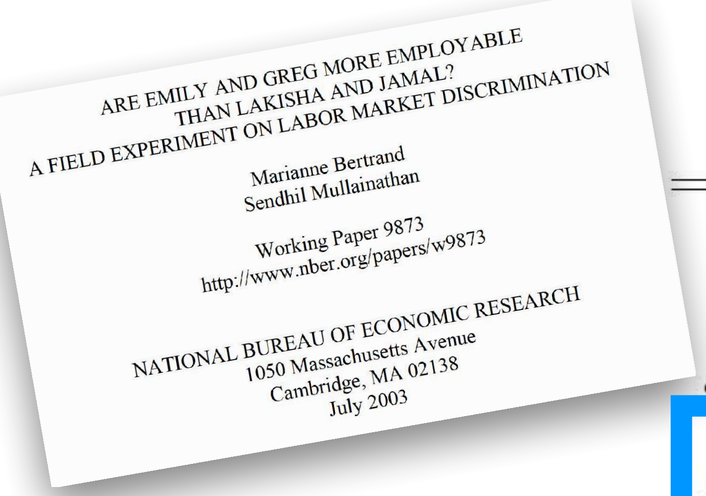


ARE EMILY AND GREG MORE EMPLOYABLE THAN LAKISHA AND JAMAL? A FIELD EXPERIMENT ON LABOR MARKET DISCRIMINATION

Marianne Bertrand
Sendhil Mullainathan

Working Paper 9873
http://www.nber.org/papers/w9873

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2003



ARE EMILY AND GREG MORE EMPLOYABLE THAN LAKISHA AND JAMAL?
A FIELD EXPERIMENT ON LABOR MARKET DISCRIMINATION

Marianne Bertrand
Sendhil Mullainathan

Working Paper 9873
http://www.nber.org/papers/w9873

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2003

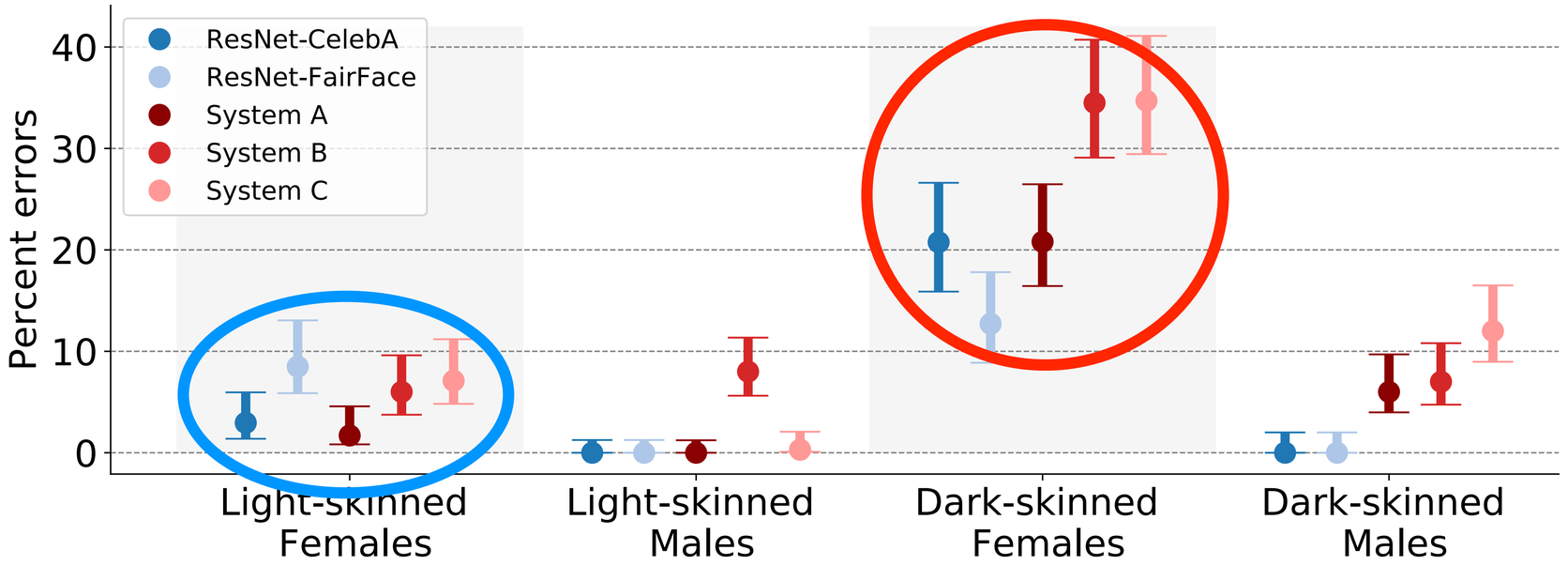
Table 1
Mean Call-Back Rates By Racial Soundingness of Names [a]

	Call-Back Rate for White Names	*Call-Back Rate for African American Names*	*Ratio*	*Difference (p-value)*
Sample:				
All sent resumes	**10.06%** [2445]	**6.70%** [2445]	**1.50**	**3.35%** (.0000)
Chicago	**8.61%** [1359]	**5.81%** [1359]	**1.48**	**2.80%** (.0024)
Boston	**11.88%** [1086]	**7.83%** [1086]	**1.52**	**4.05%** (.0008)
Females	**10.33%** [1868]	**6.87%** [1893]	**1.50**	**3.46%** (.0001)
Females in administrative jobs	**10.93%** [1363]	**6.81%** [1364]	**1.60**	**4.12%** (.0001)
Females in sales jobs	**8.71%** [505]	**6.99%** [529]	**1.25**	**1.72%** (.1520)
Males	**9.19%** [577]	**6.16%** [552]	**1.49**	**3.03%** (.0283)

Observation -> correlation

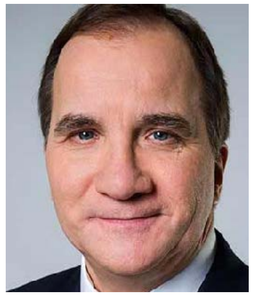
Experiment -> cause

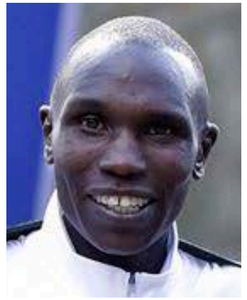
Gender classification



PPB dataset
(Scandinavia + Sub-Saharan Africa)
[Buolamwini & Gebru 2018]

Gender-typical

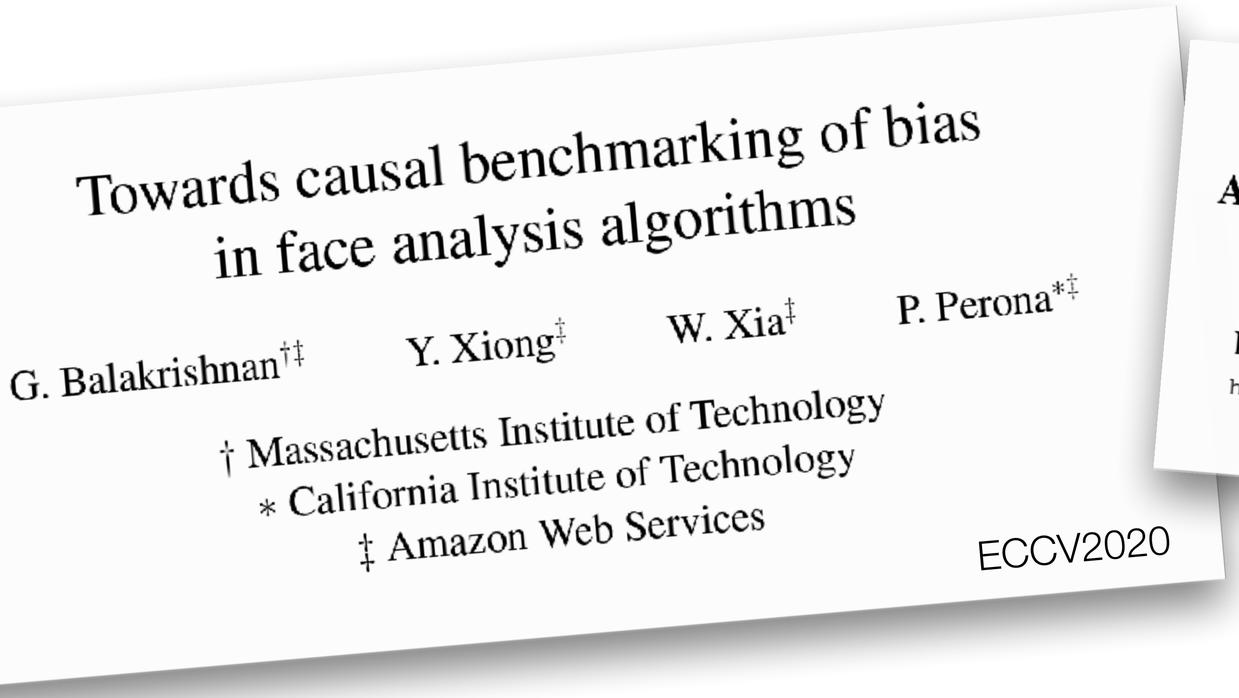


Non-typical




[Mutukhumar et al. 2018]

Experimental method in computer vision?


Towards causal benchmarking of bias in face analysis algorithms

G. Balakrishnan[†‡] Y. Xiong[‡] W. Xia[‡] P. Perona[*‡]

† Massachusetts Institute of Technology
∗ California Institute of Technology
‡ Amazon Web Services

ECCV2020


Benchmarking Algorithmic Bias in Face Recognition: An Experimental Approach Using Synthetic Faces and Human Evaluation

Hao Liang
Rice University
hl106@rice.edu

Pietro Perona
California Institute of Technology and AWS
perona@caltech.edu, peronapp@amazon.com

Guha Balakrishnan
Rice University
guha@rice.edu

ICCV2023

Darker skin

Idea 1: use synthetic faces

https://www.thispersondoesnotexist.com/





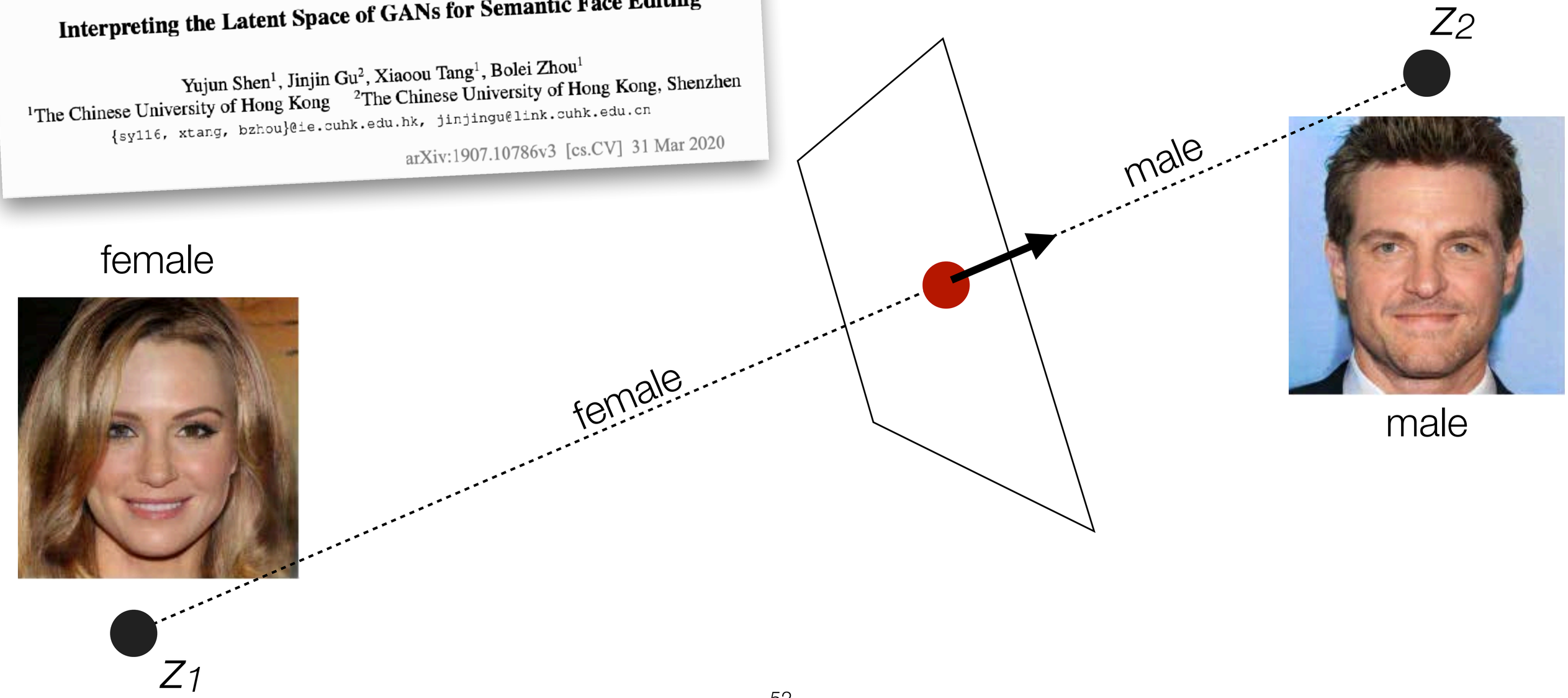



Idea 2: Control of latent space

Interpreting the Latent Space of GANs for Semantic Face Editing

Yujun Shen[1], Jinjin Gu[2], Xiaoou Tang[1], Bolei Zhou[1]
[1]The Chinese University of Hong Kong [2]The Chinese University of Hong Kong, Shenzhen
{sy116, xtang, bzhou}@ie.cuhk.edu.hk, jinjingu@link.cuhk.edu.cn

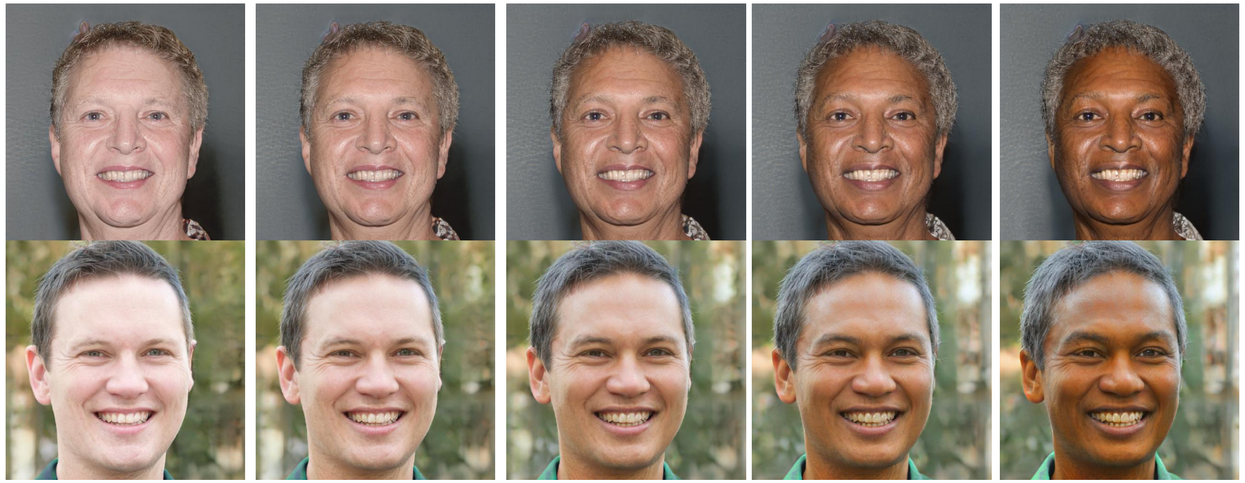
arXiv:1907.10786v3 [cs.CV] 31 Mar 2020



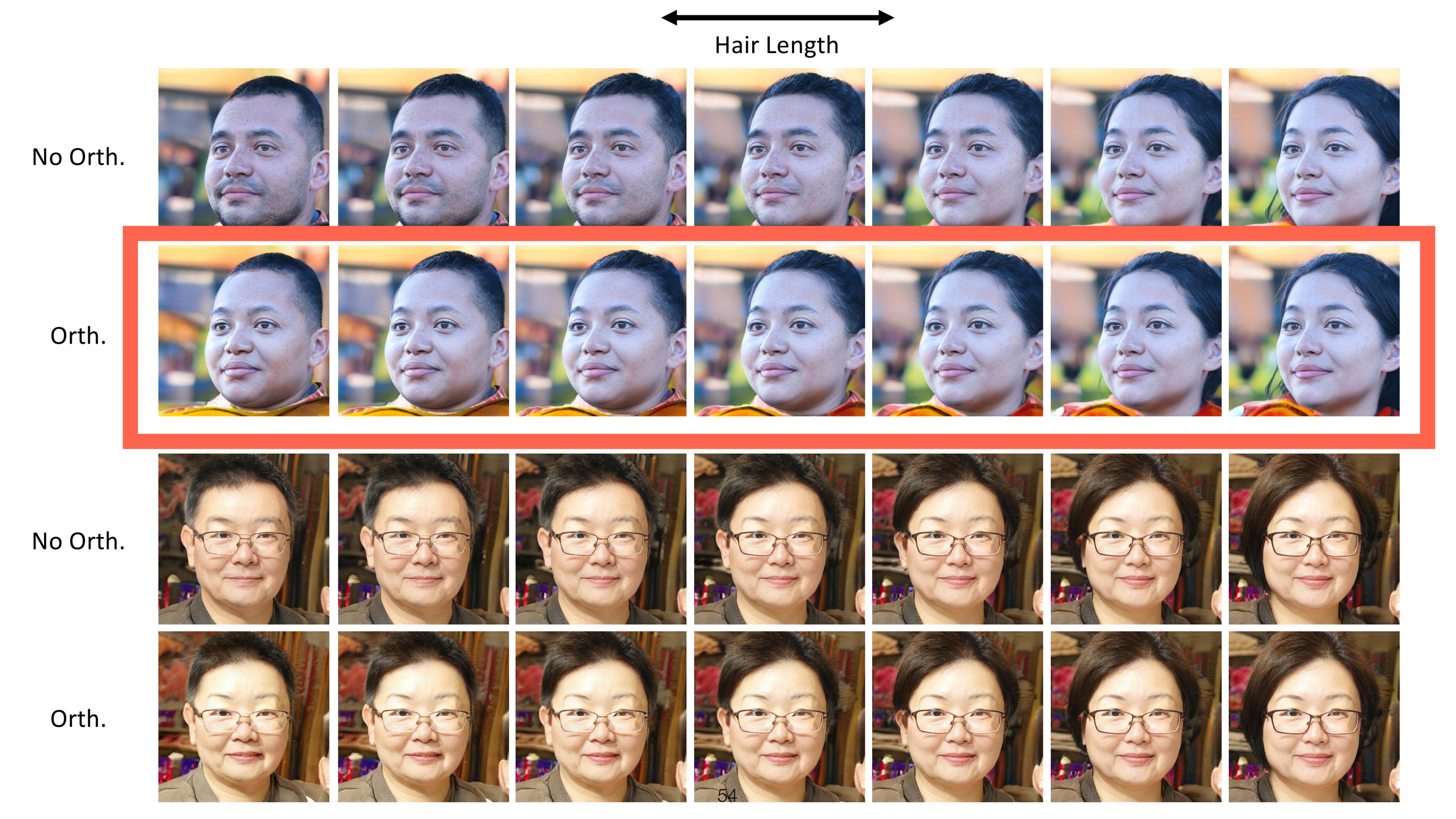
Age

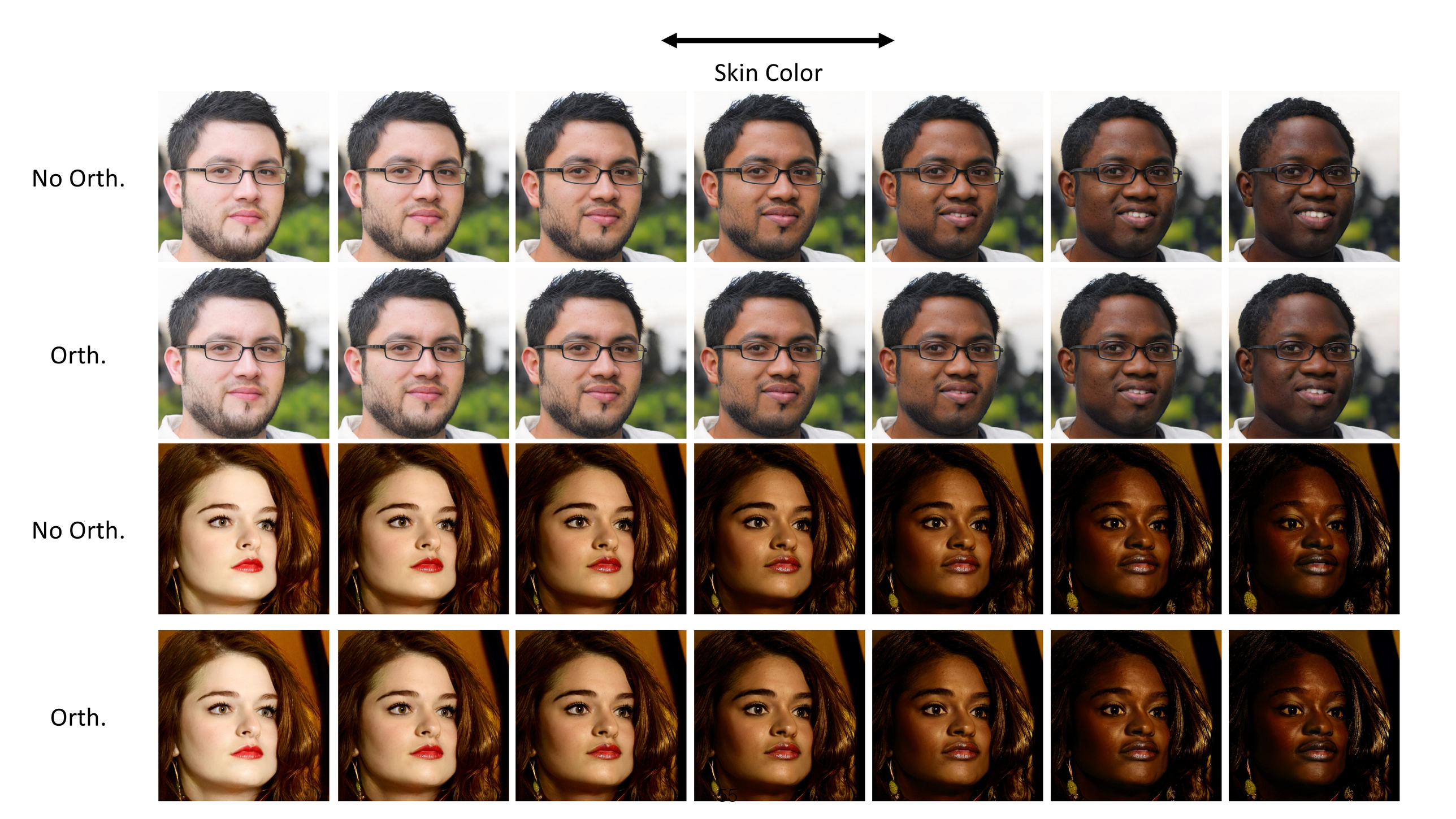
Facial Hair

Skin Color

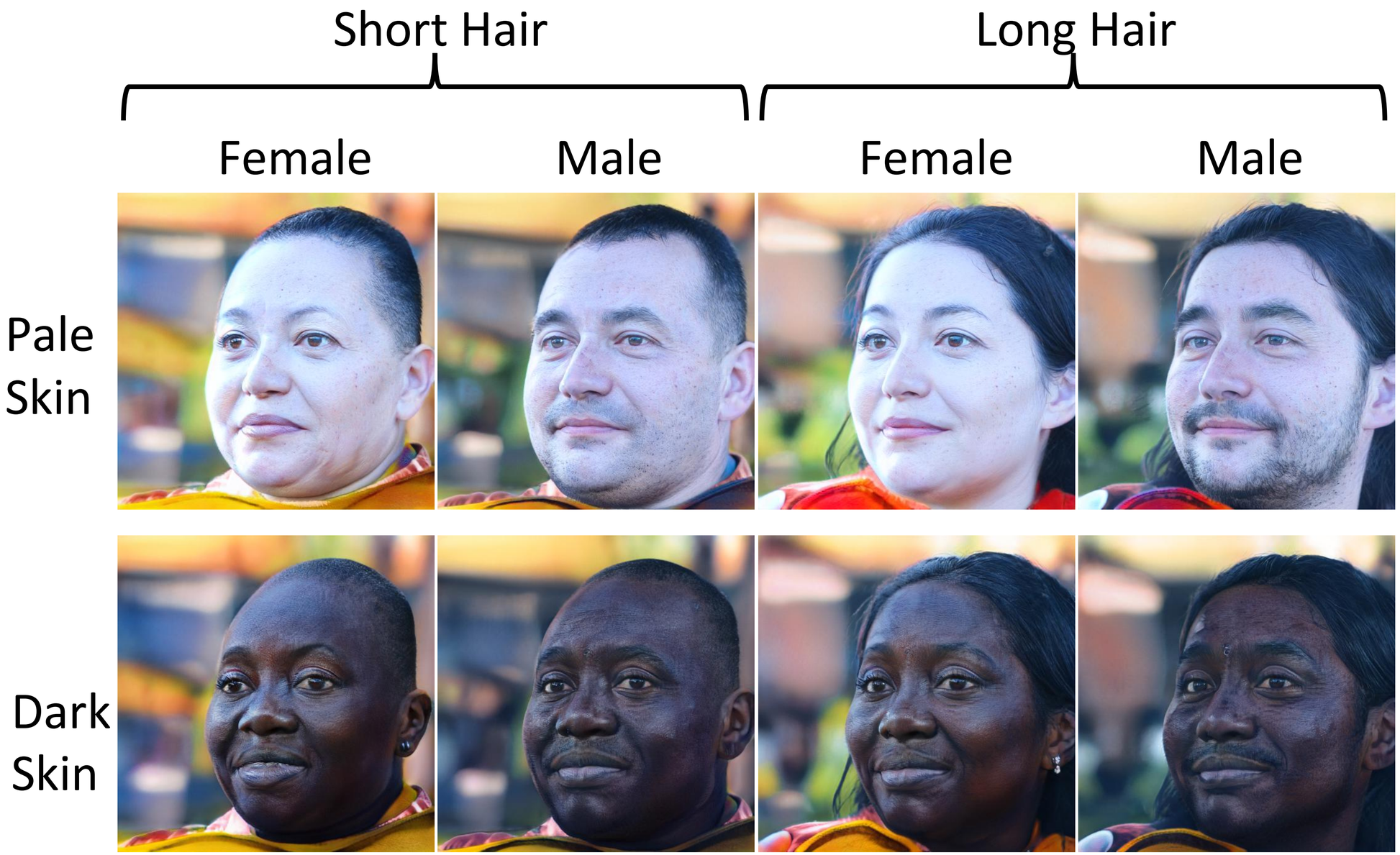


Hair Length

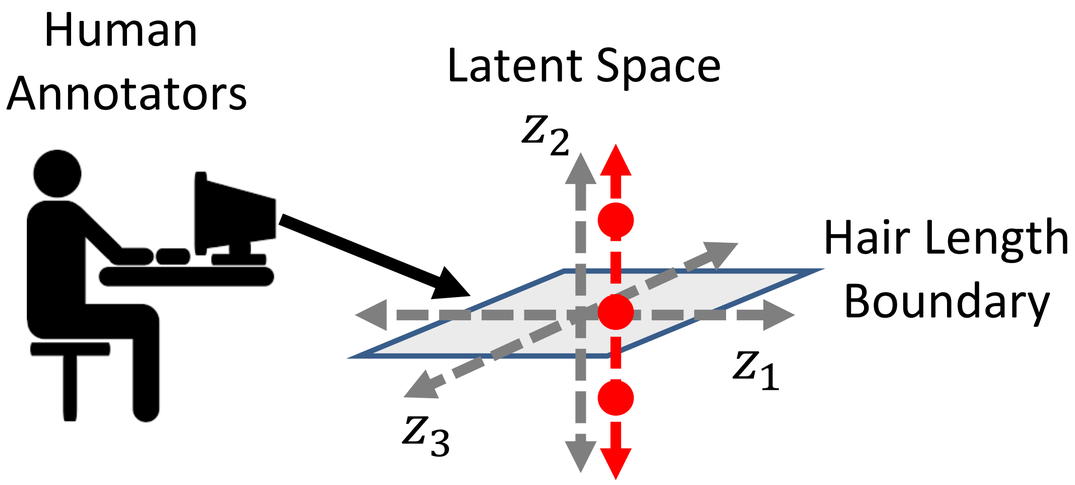

Hair Length
No Orth.
Orth.
No Orth.
Orth.

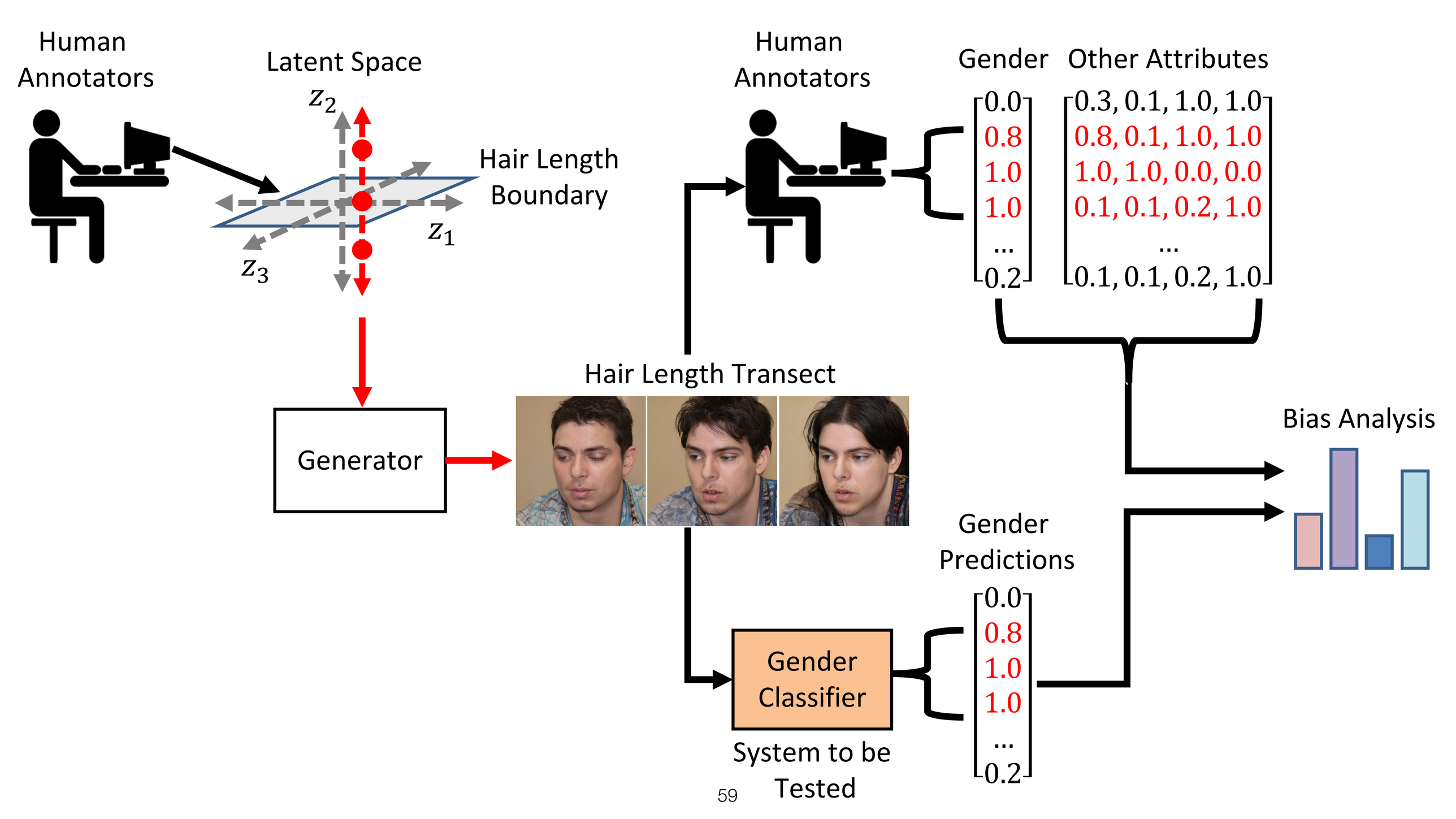

Skin Color
No Orth.
Orth.
No Orth.
Orth.

Idea 3: matched samples



Idea 4: calibrate with human annotators


Human
Annotators
Latent Space
z_2
Hair Length
Boundary
z_1
z_3


Human Annotators
Latent Space
z_2
z_1
z_3
Hair Length Boundary
Generator
Hair Length Transect
Human Annotators
Gender
Other Attributes
0.0
0.8
1.0
1.0
...
0.2
0.3, 0.1, 1.0, 1.0
0.8, 0.1, 1.0, 1.0
1.0, 1.0, 0.0, 0.0
0.1, 0.1, 0.2, 1.0
...
0.1, 0.1, 0.2, 1.0
Bias Analysis
Gender Predictions
Gender Classifier
System to be Tested

Collecting and assessing human annotations

Instructions

View full instructions

View tool guide

NOTE: Please do not worry if the face looks a little weird!

What's the skin color of this person? (Please refer to the scale on the left.)

Select an option

Light	1
Fair	2
Medium	3
Olive	4
Brown	5
Black	6

Instructions

View full instructions

View tool guide

* **Very short**: Bald or buzz cut. Between 0 and 1" (0-2.5cm)

* **Short**: 1-2" (2.5-5cm). Not easy to comb.

* **Medium**: May be combed, ears partially covered. 2-4" (5-10cm).

***Long**: Covers ears, may reach shoulders. 4-8" (10-20cm).

***Very long**. Beyond shoulder length.

What's the length of the hair of the person in this photograph?

Select an option

Very short	1
Short	2
Medium	3
Long	4
Very long	5

SageMaker Ground Truth

Easily label training data for machine learning at scale

Instructions

View full instructions

View tool guide

* **Female**: Everyone will agree.

* **Probably female**: Most will say female. A few may say male.

***In between**: Could be male or female or gender-neutral.

***Probably male**: Most will say male. A few might say female.

***Male**. Everyone will agree.

NOTE: Please do not worry if the face looks a bit weird!

What would casual observers think about the gender of this face?

Select an option

Female	1
Probably female	2
In between	3
Probably male	4
Male	5

Zoom in Zoom out Move Fit image

Submit

Instructions

View full instructions

View tool guide

* **Child**: Between 0 and 12. Before puberty.

* **Teen**: 13-18, in middle/high school.

* **Young adult**: 19-30. Attending college, first job.

* **Adult:** 30-50. Parent, mid-career.

***Middle age**: 50-65 - Late career, kids in college, approaching retirement.

* **Senior**. 65+. Often retired, may be elderly.

NOTE: Please do not worry if the face looks a bit weird.

What's the approximate age of this person?

Select an option

Child	1
Teen	2
Young adult	3
Adult	4
Middle age	5
Senior	6

Zoom in Zoom out Move Fit image

Submit

Instructions

View full instructions

View tool guide

* **Real for sure**: Most people would agree that this is real.

* **Likely real**: Many people would think it's real, but some would have doubts.

* **Uncertain:** People could come out either way.

***Likely fake:** Many would say it's fake, but not everyone.

***Fake:** Most people would not think that this is real.

Is this the photograph of a real face, or a fake generated by a computer?

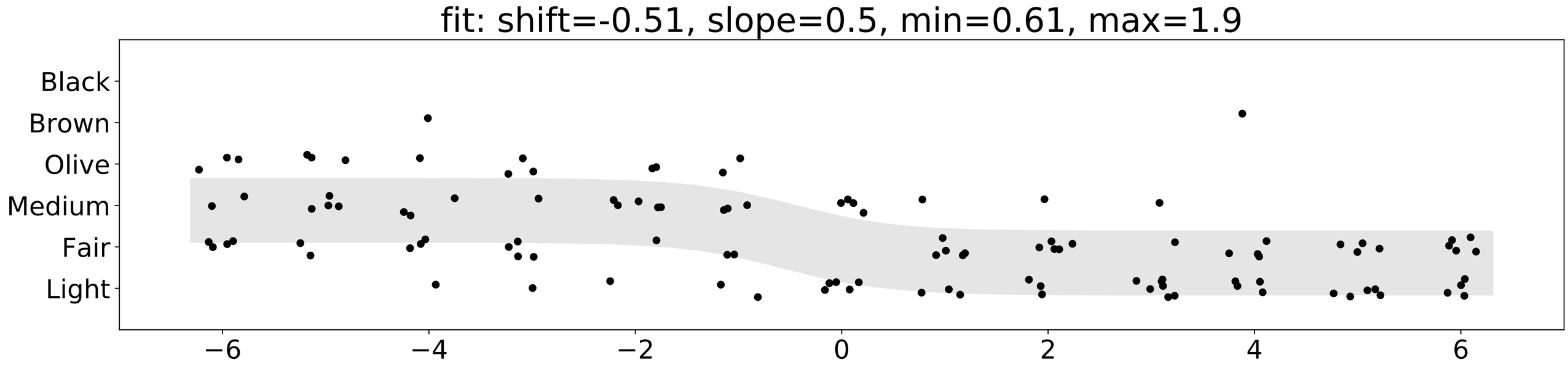
Select an option

Real for sure	1
Likely real	2
Uncertain	3
Likely fake	4
Fake for sure	5

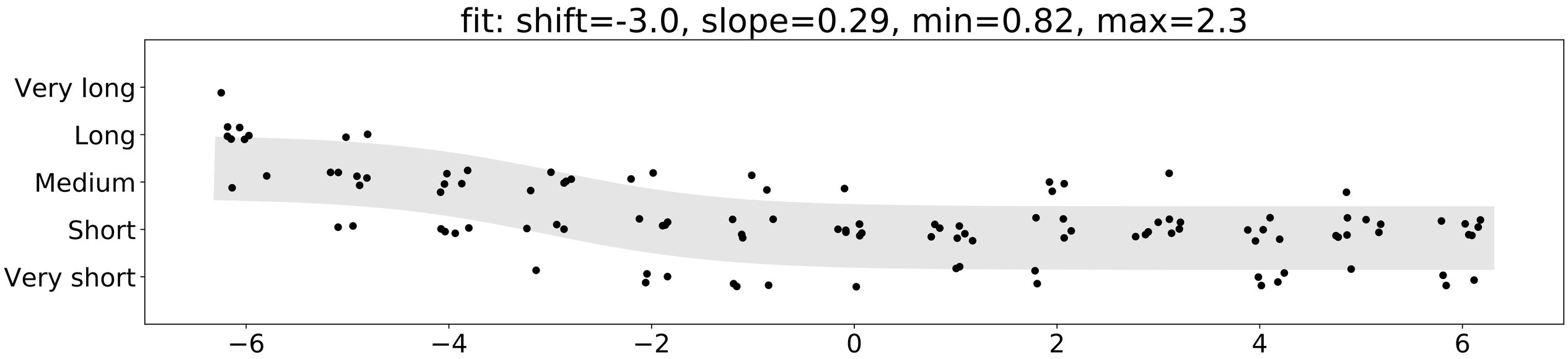
Zoom in Zoom out Move Fit image

Submit

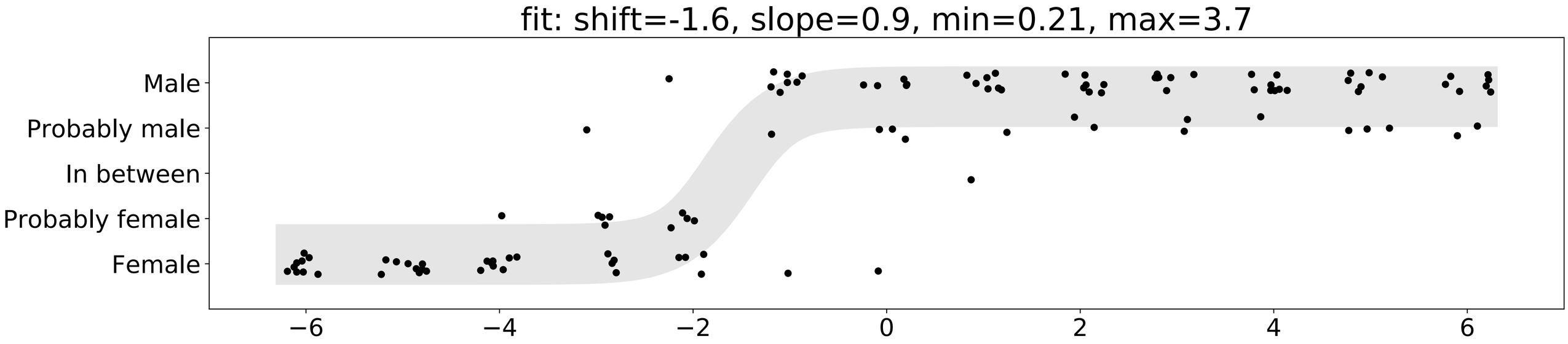
Skin tone

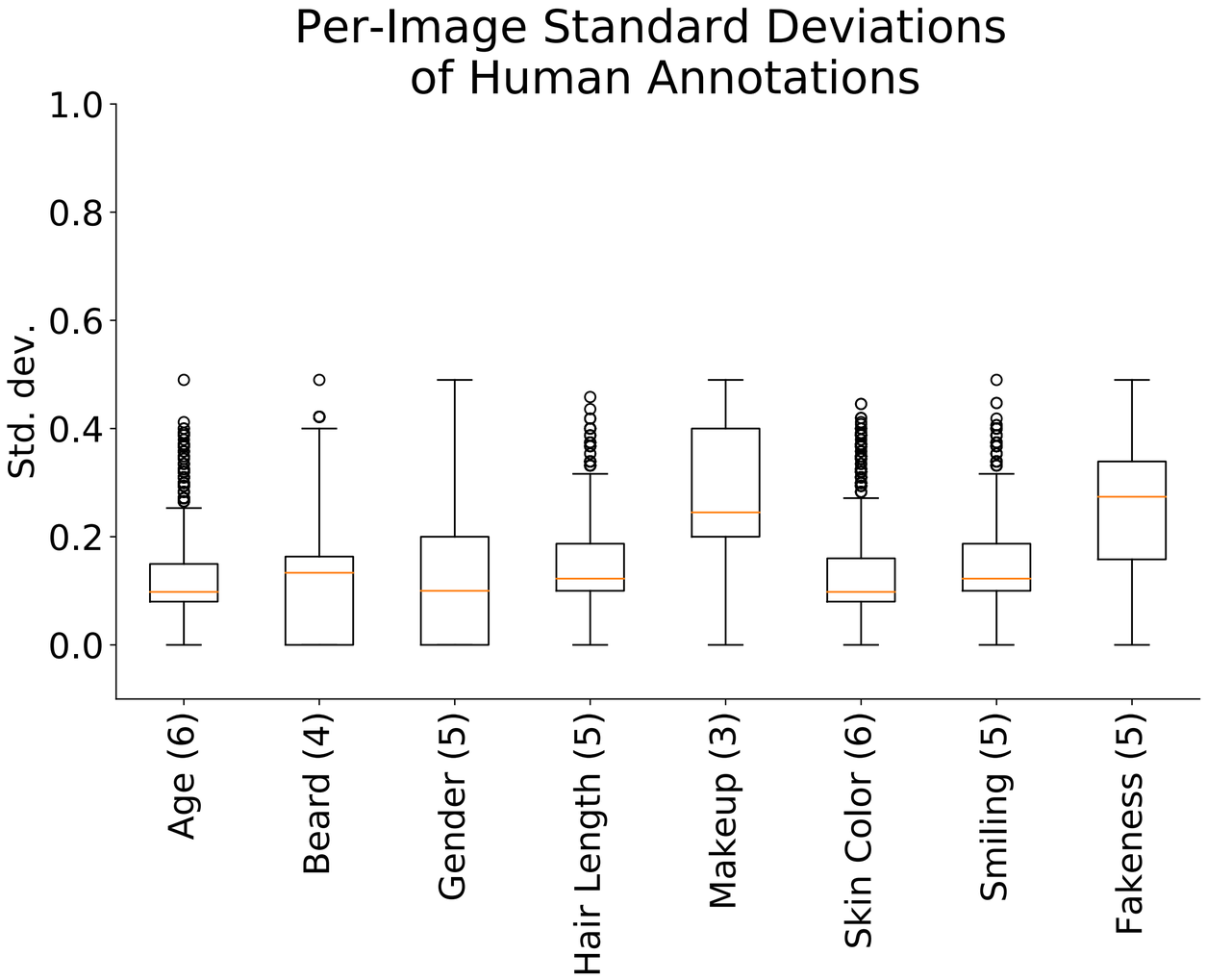

fit: shift=-0.51, slope=0.5, min=0.61, max=1.9
Black
Brown
Olive
Medium
Fair
Light
−6
−4
−2
0
2
4
6

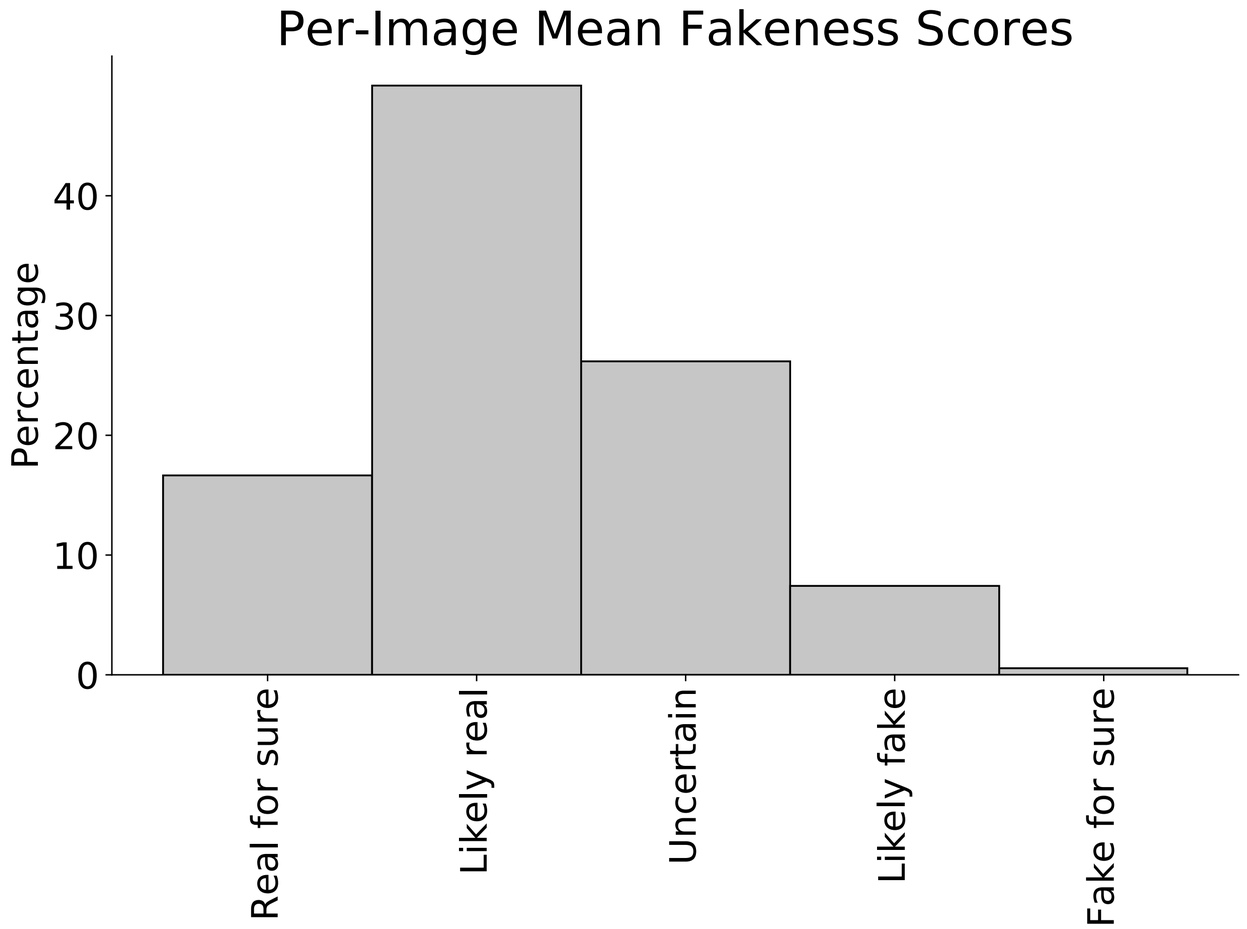
Hair length

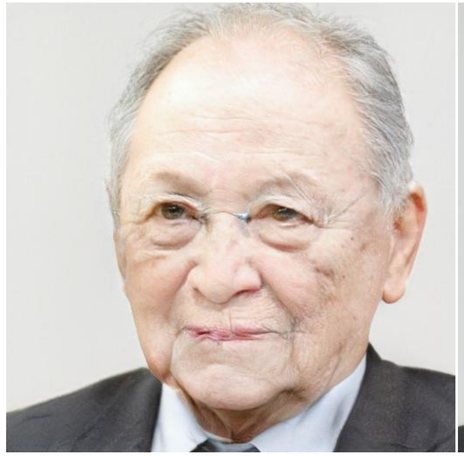

fit: shift=-3.0, slope=0.29, min=0.82, max=2.3
Very long
Long
Medium
Short
Very short
−6
−4
−2
0
2
4
6

Gender

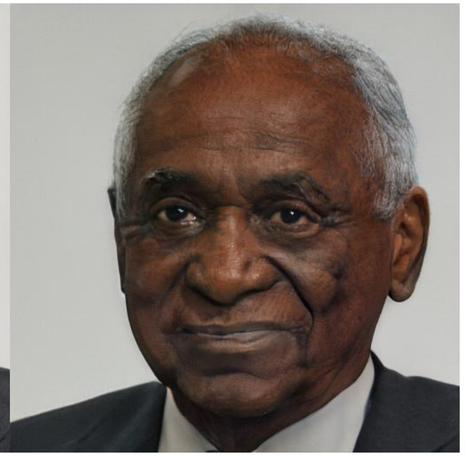

fit: shift=-1.6, slope=0.9, min=0.21, max=3.7
Male
Probably male
In between
Probably female
Female
−6
−4
−2
0
2
4
6


Per-Image Standard Deviations
of Human Annotations
Std. dev.
1.0
0.8
0.6
0.4
0.2
0.0
Age (6)
Beard (4)
Gender (5)
Hair Length (5)
Makeup (3)
Skin Color (6)
Smiling (5)
Fakeness (5)


Per-Image Mean Fakeness Scores
Percentage
40
30
20
10
0
Real for sure
Likely real
Uncertain
Likely fake
Fake for sure




Humans:	Male	Male	Likely Female	Male
Intended:	Female	Male	Female	Male

	Likely Female	Male	Female	Likely Male
	Female	Male	Female	Male

Hair length

0-0.1 | 0.1-0.2 | 0.2-0.3 | **0.3-0.4** | **0.4-0.5** | 0.5-0.6 | 0.6-0.7 | 0.7-0.8 | 0.8-0.9 | 0.9-1

Gender

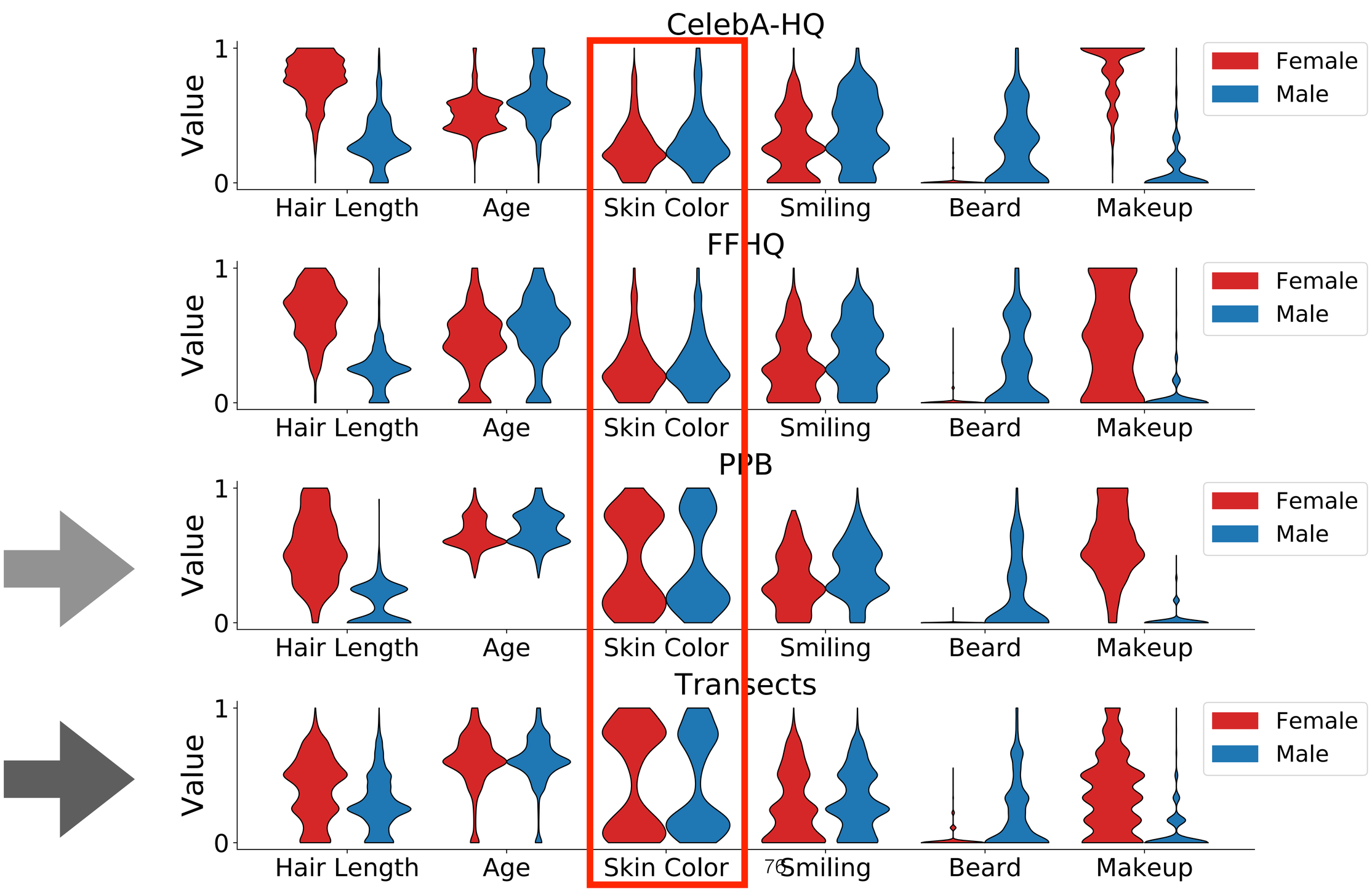
0-0.1	0.1-0.2	0.2-0.3	0.3-0.4	0.4-0.5	0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1

Skin tone

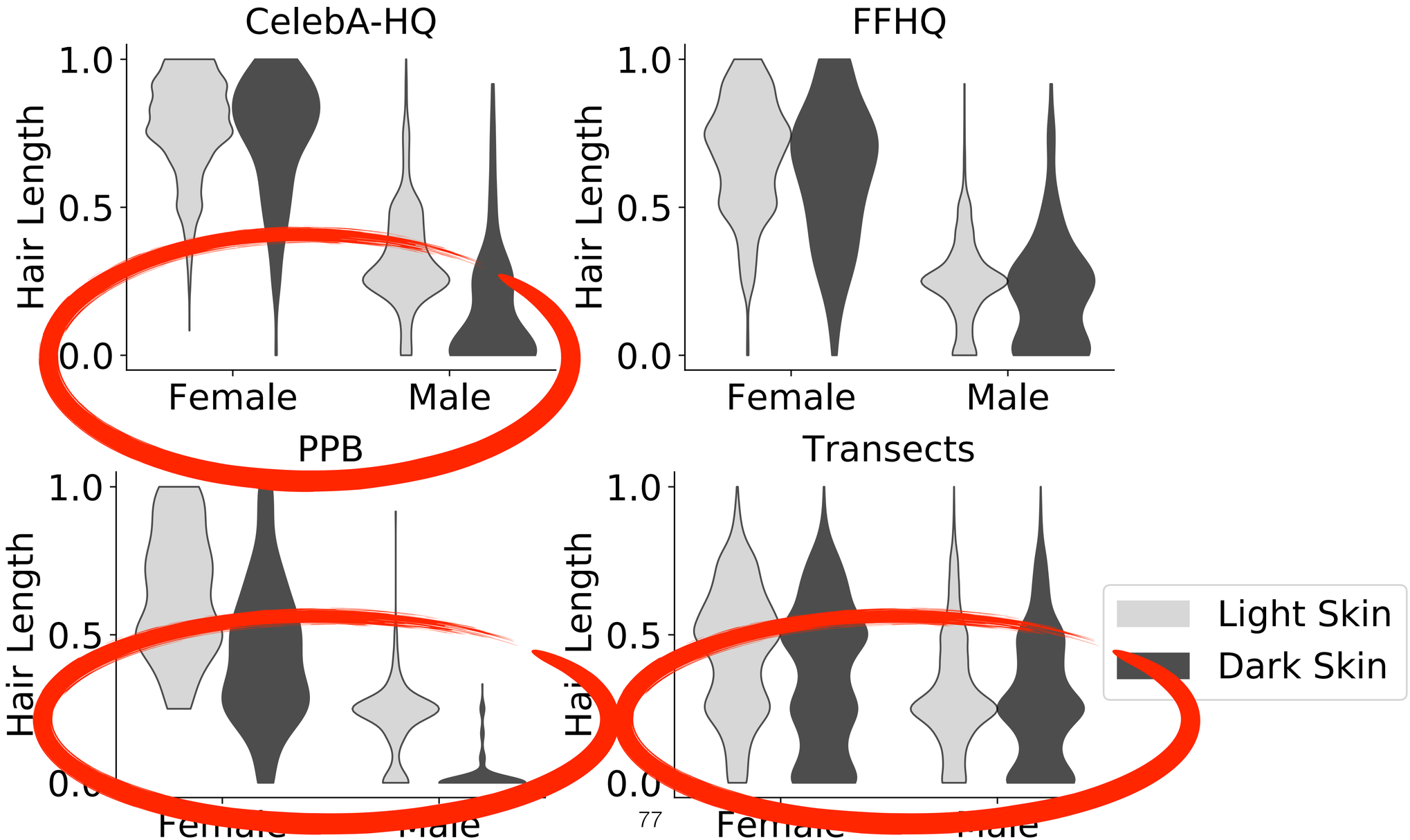
Gender

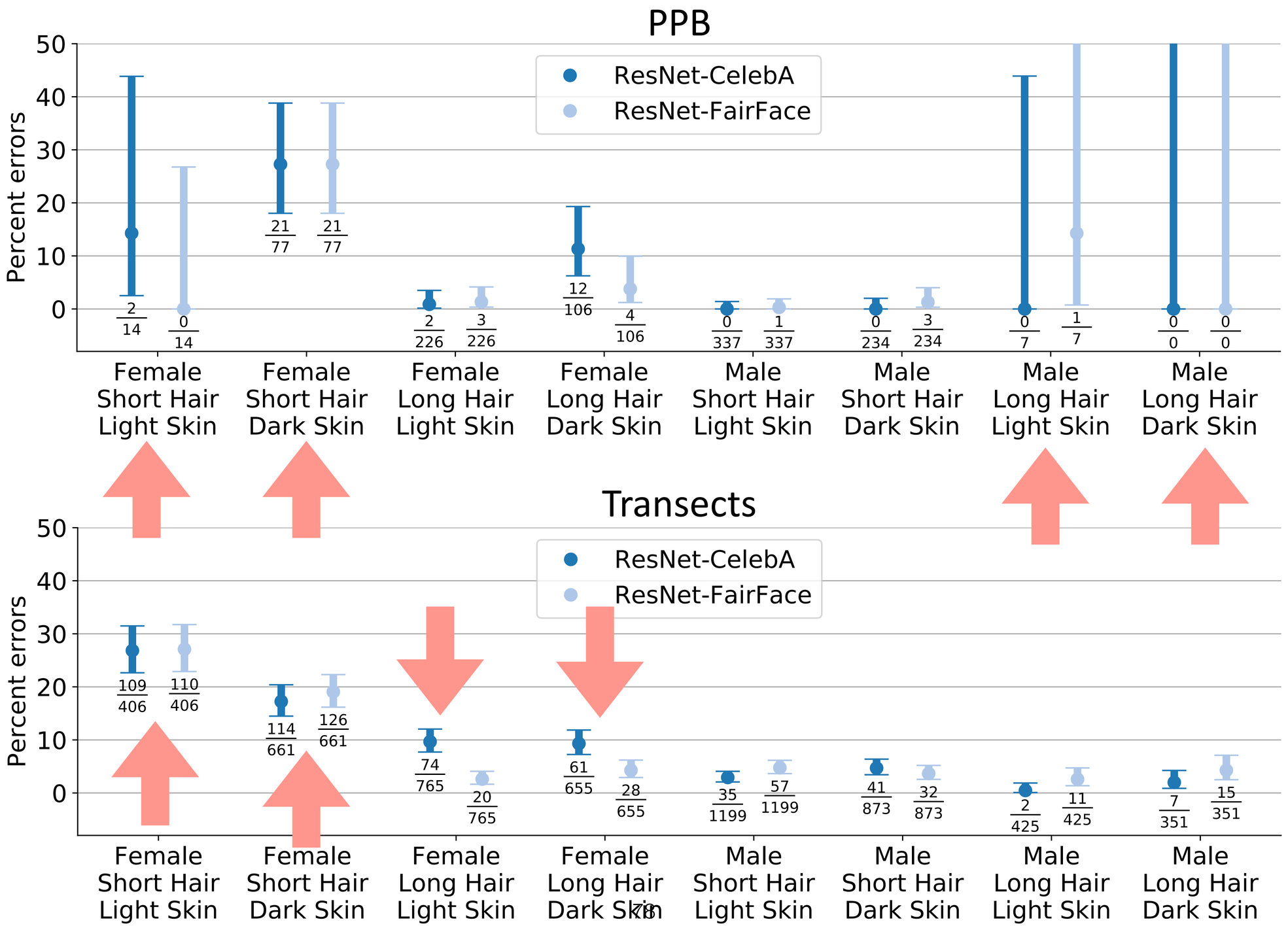
< 0.2 | 0.2-0.3 | 0.3-0.4 | **0.4-0.5** | **0.5-0.6** | 0.6-0.7 | 0.7-0.8 | > 0.8

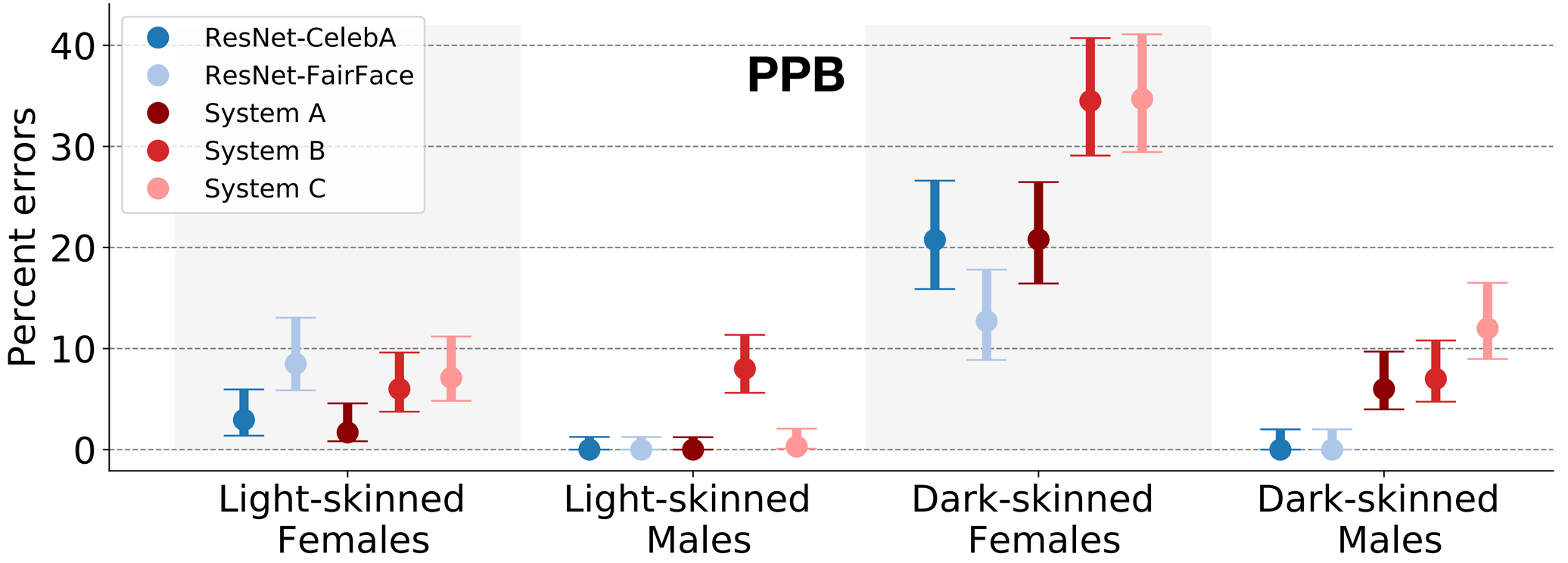
Data statistics and analysis

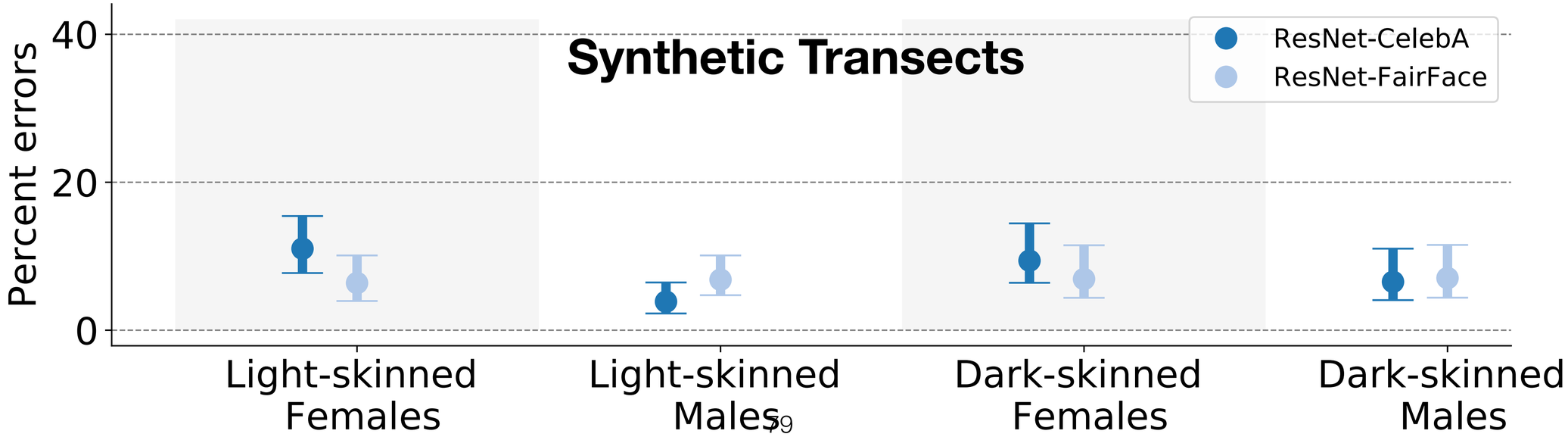

CelebA-HQ
FFHQ
PPB
Transects
Value
1
0
Hair Length
Age
Skin Color
Smiling
Beard
Makeup
Female
Male

Intersectional

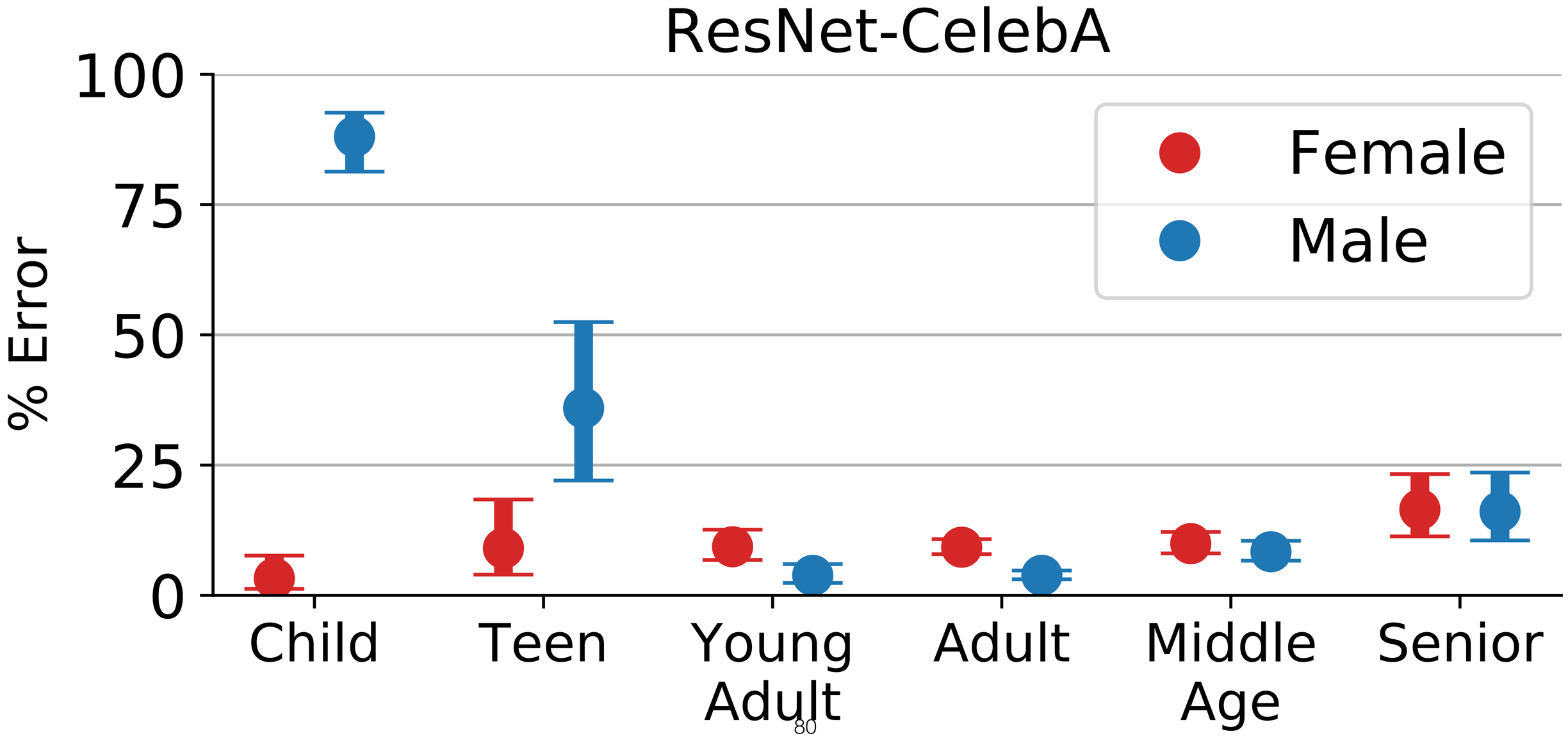



PPB
ResNet-CelebA
ResNet-FairFace
Percent errors
0
10
20
30
40
50
2/14
0/14
21/77
21/77
2/226
3/226
12/106
4/106
0/337
1/337
0/234
3/234
0/7
1/7
0/0
0/0
Female Short Hair Light Skin
Female Short Hair Dark Skin
Female Long Hair Light Skin
Female Long Hair Dark Skin
Male Short Hair Light Skin
Male Short Hair Dark Skin
Male Long Hair Light Skin
Male Long Hair Dark Skin
Transects
ResNet-CelebA
ResNet-FairFace
Percent errors
0
10
20
30
40
50
109/406
110/406
114/661
126/661
74/765
20/765
61/655
28/655
35/1199
57/1199
41/873
32/873
2/425
11/425
7/351
15/351
Female Short Hair Light Skin
Female Short Hair Dark Skin
Female Long Hair Light Skin
Female Long Hair Dark Skin
Male Short Hair Light Skin
Male Short Hair Dark Skin
Male Long Hair Light Skin
Male Long Hair Dark Skin


PPB
ResNet-CelebA
ResNet-FairFace
System A
System B
System C
Percent errors
0
10
20
30
40
Light-skinned Females
Light-skinned Males
Dark-skinned Females
Dark-skinned Males


Synthetic Transects
ResNet-CelebA
ResNet-FairFace
Percent errors
0
20
40
Light-skinned Females
Light-skinned Males
Dark-skinned Females
Dark-skinned Males

Age and gender

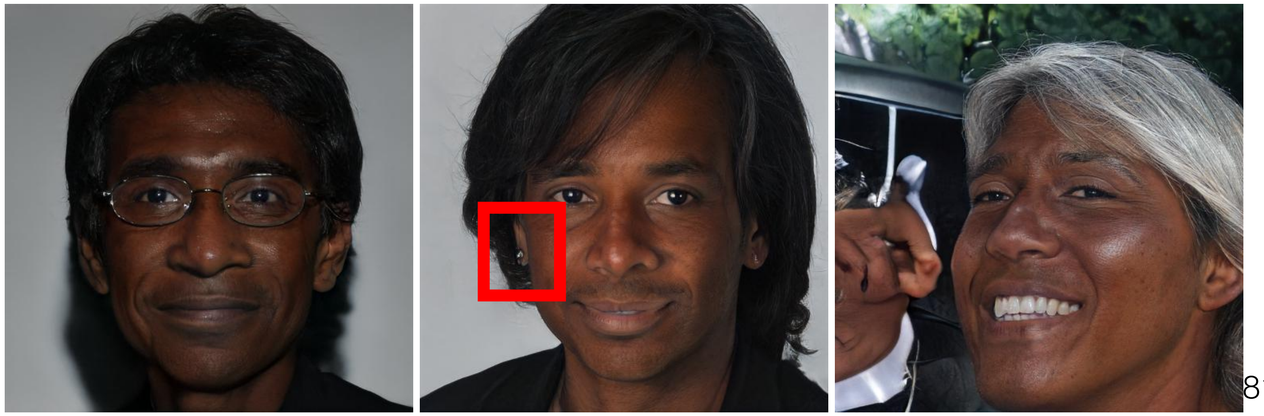


Residual correlations

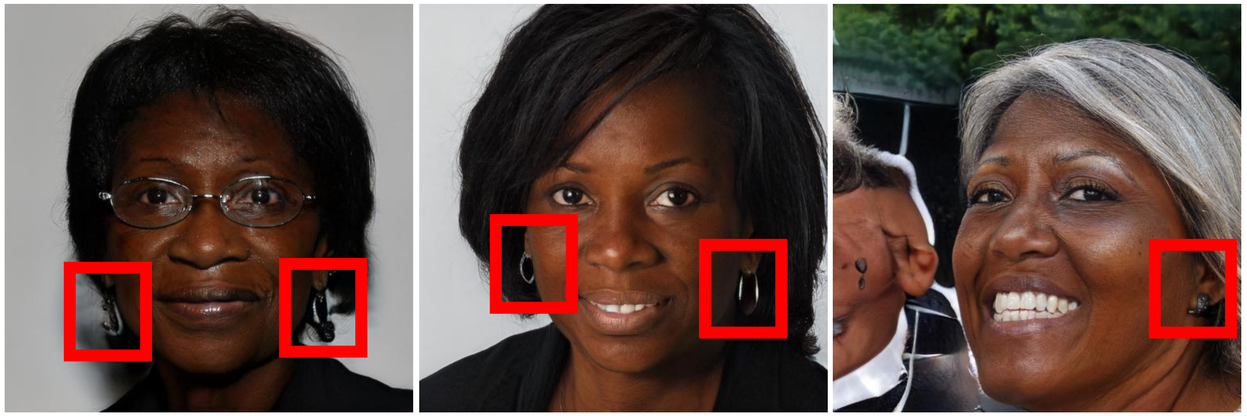
Example 1

Example 2

Males



Females



Quiz

Race Bias

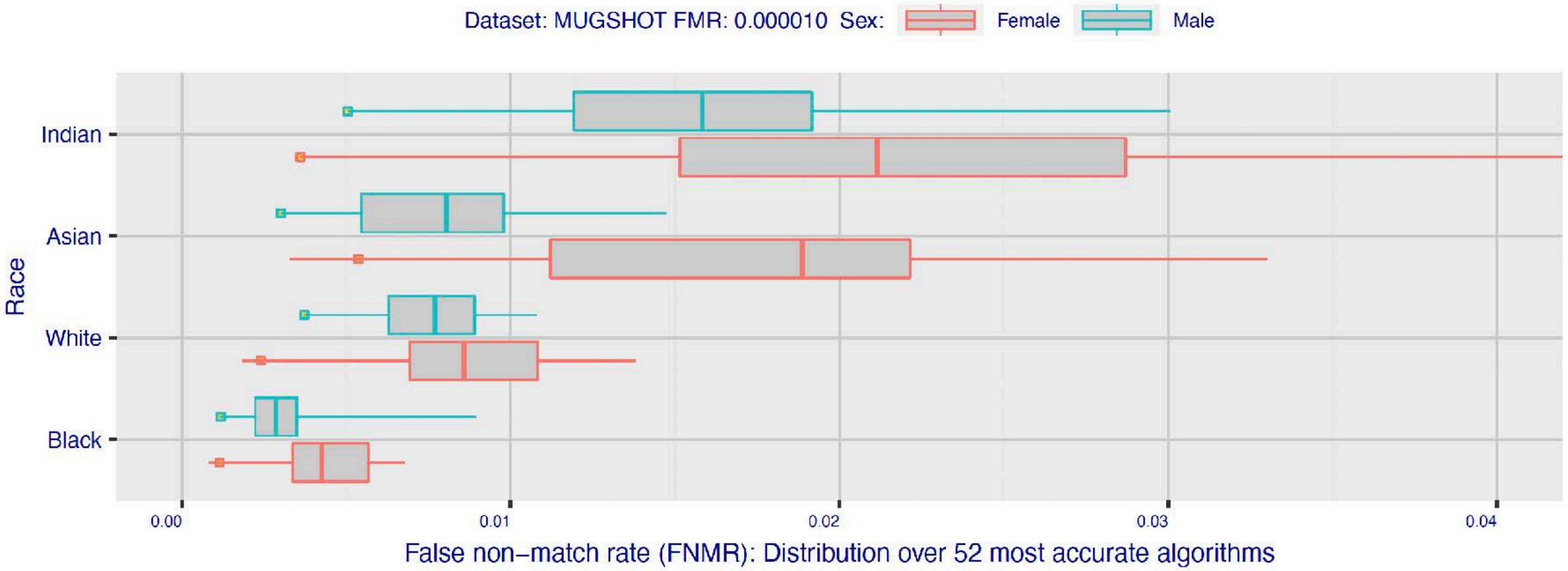


Figure 17: For mugshot comparisons, the figure shows the distribution of FNMR values over the 52 most accurate verification algorithms, by sex and race. The threshold was set for each algorithm to achieve FMR = 0.00001 over all imposter comparisons. The line within each box is the median over those algorithms; the box itself spans the interquartile range (26 algorithms) and the lines here extend to minimum and maximum values. The small box on the left side indicates the accuracy for best algorithm overall, on this dataset alphaface001.

Race bias

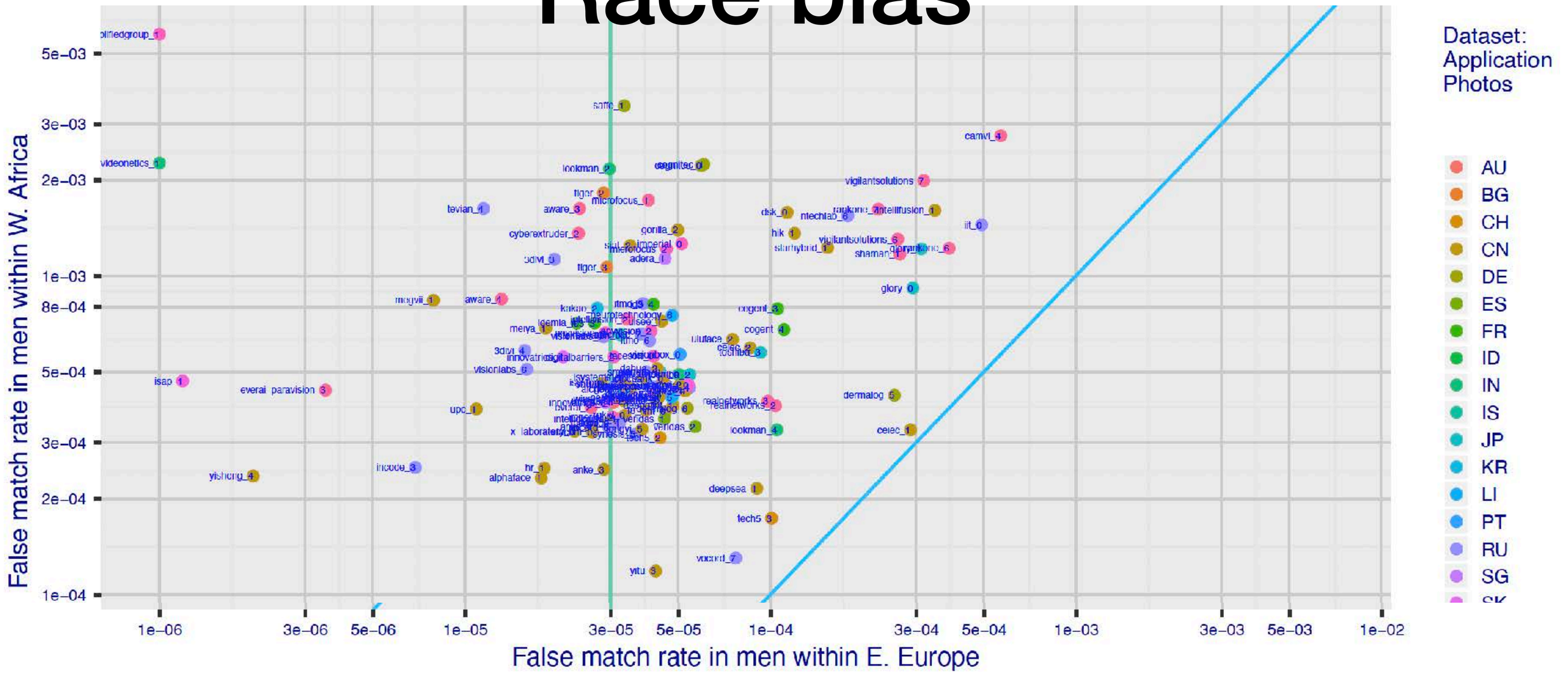
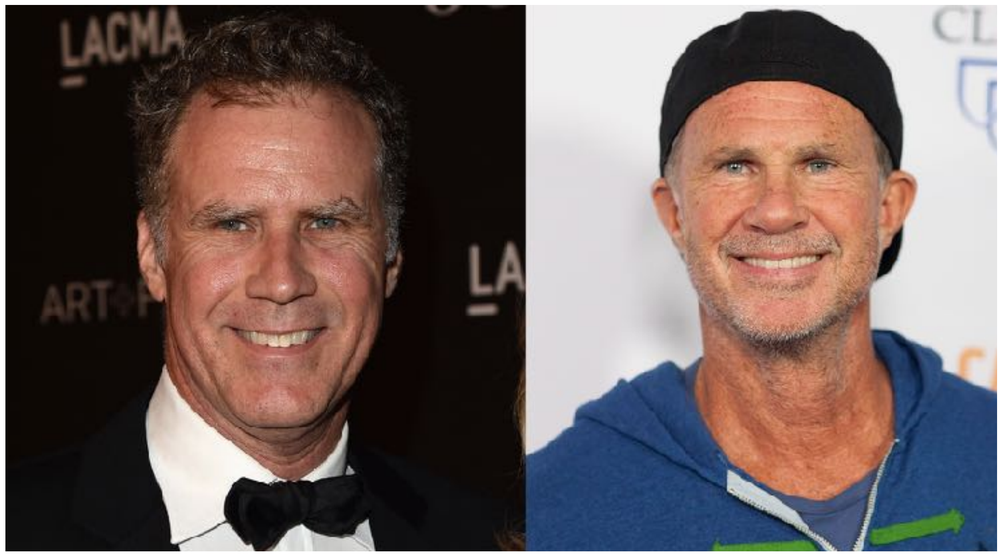
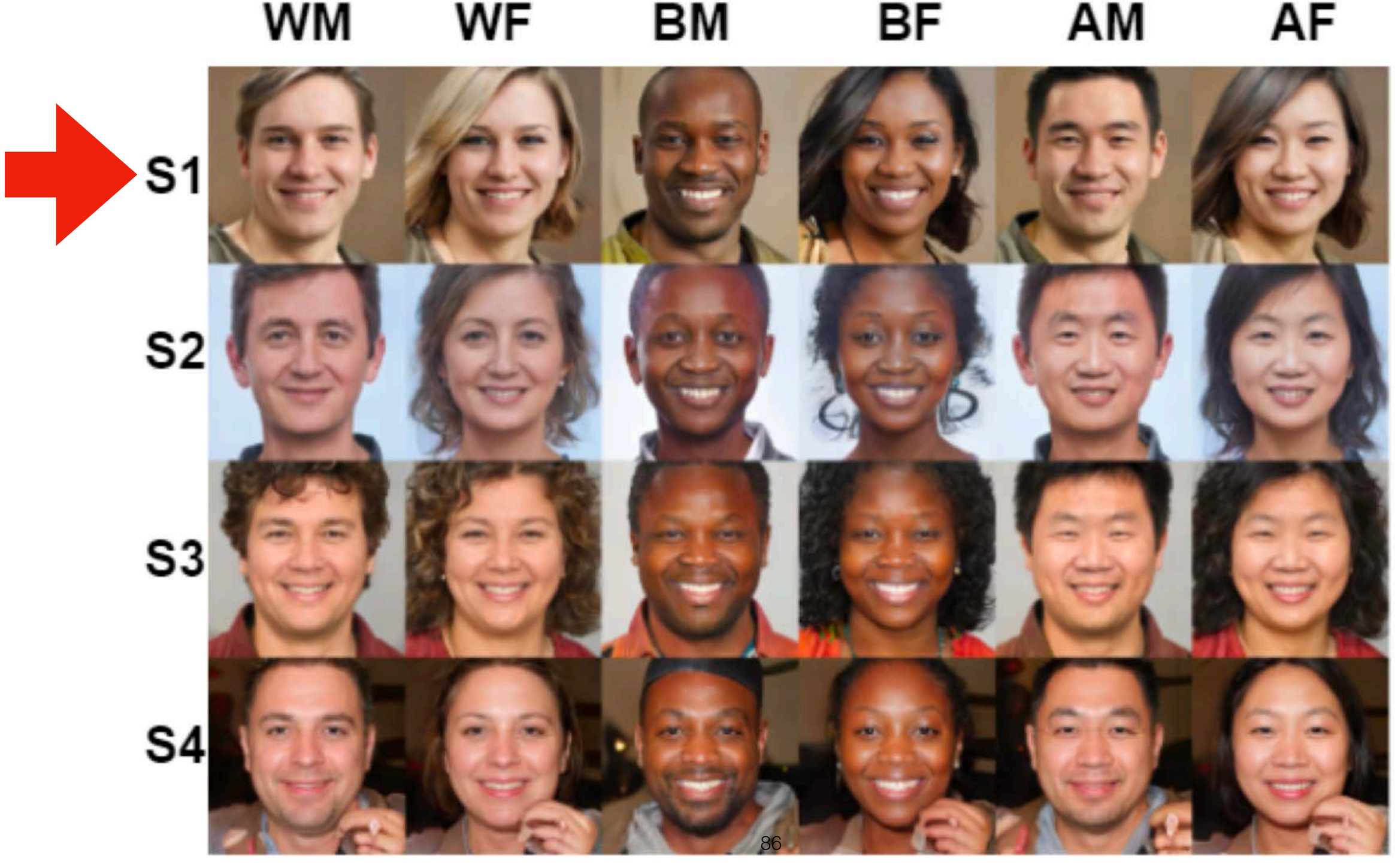


Figure 8: The scatter plot shows FMR when comparing same-age **men** *within and across three Eastern European count (Russia, Ukraine, Poland), against FMR obtained comparing men within and across three West African countries (Gh Liberia, Nigeria). The threshold is fixed for each algorithm to give the FMR noted in the annotation over white me the U.S. mugshot database. This is indicated by the vertical and horizontal green lines. The blue diagonal line $y =$*

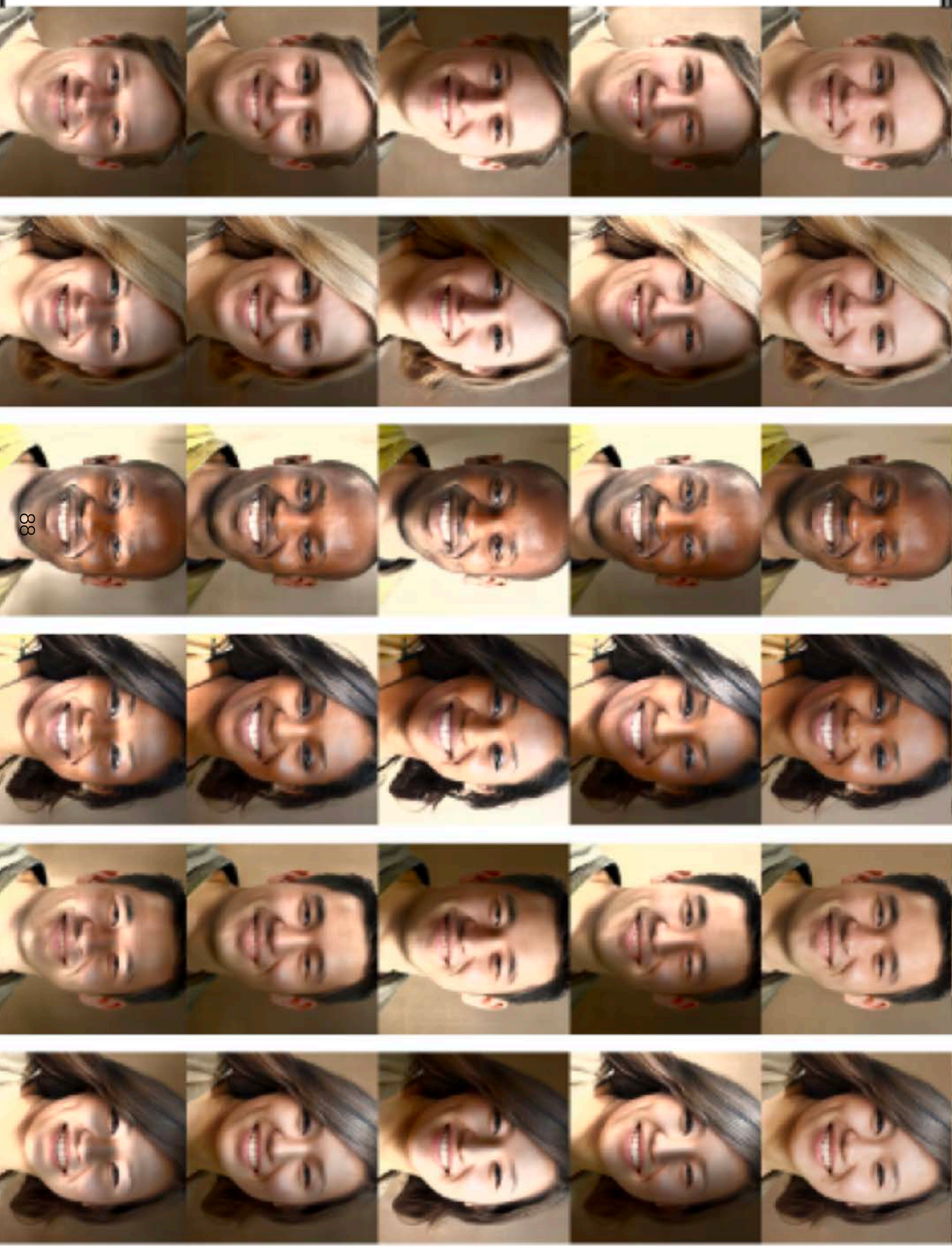
Identity




WM
WF
BM
BF
AM
AF
S1
S2
S3
S4

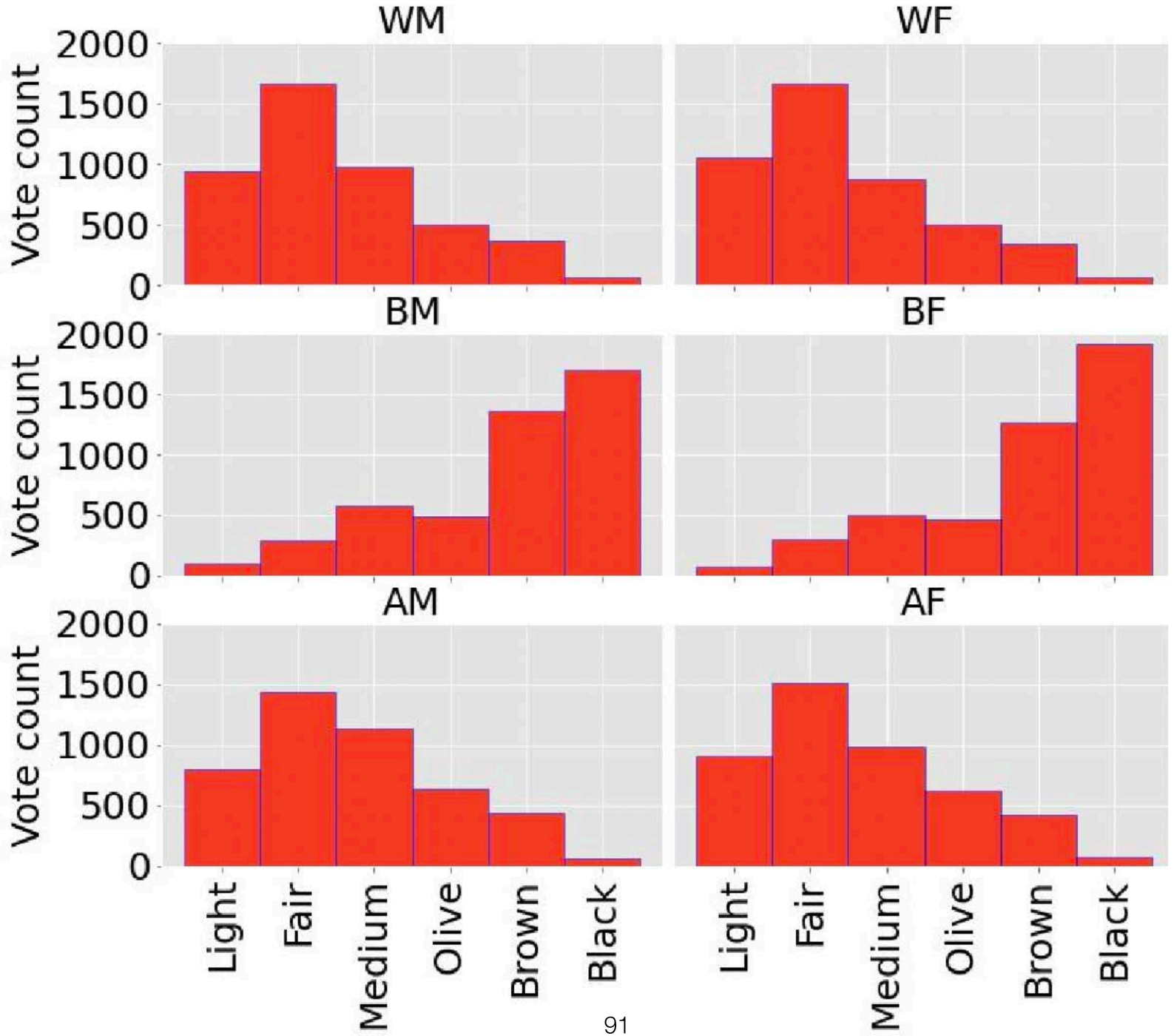

Pose

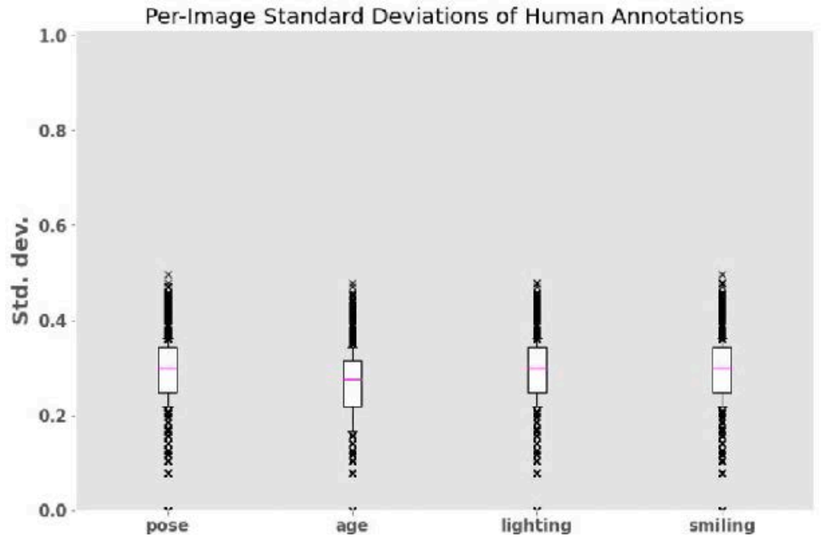
Lighting



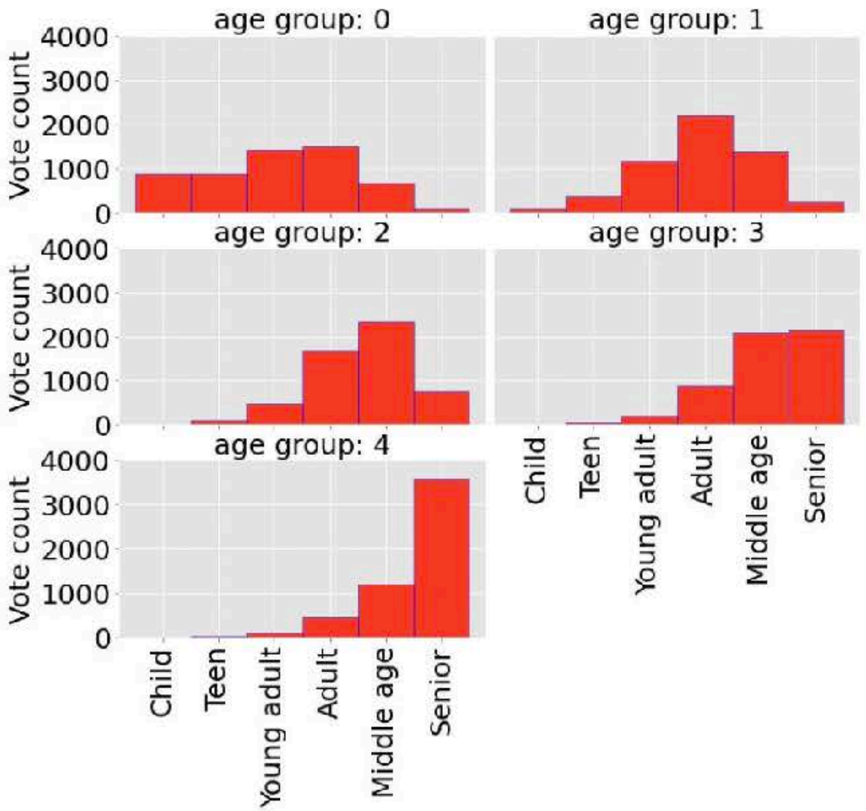
Age

Expression

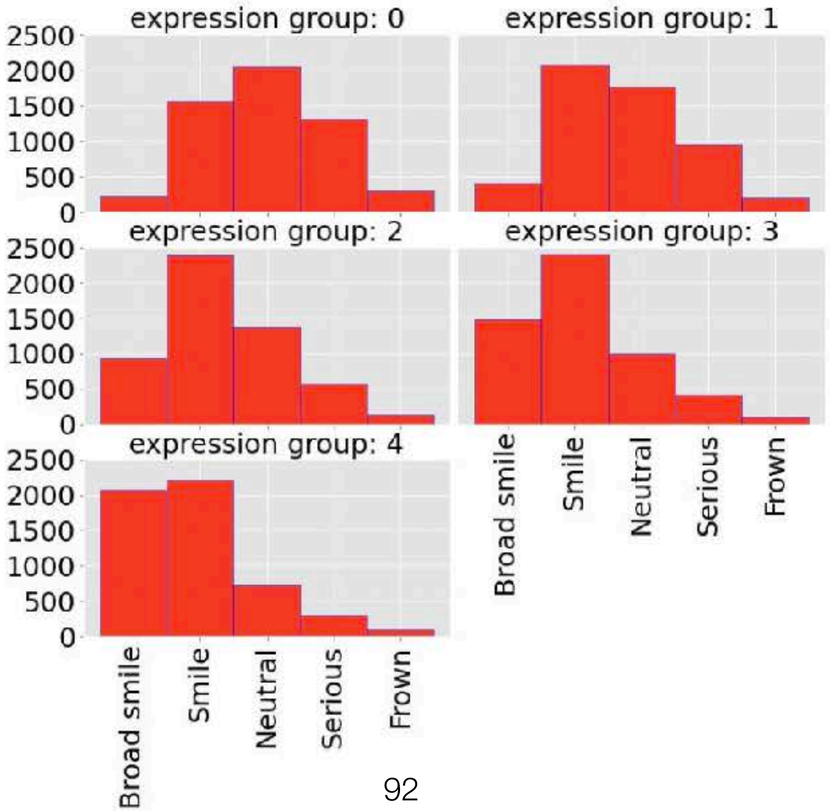

WM
WF
BM
BF
AM
AF
Vote count
2000
1500
1000
500
0
Light
Fair
Medium
Olive
Brown
Black


Per-Image Standard Deviations of Human Annotations
Std. dev.
1.0
0.8
0.6
0.4
0.2
0.0
pose
age
lighting
smiling

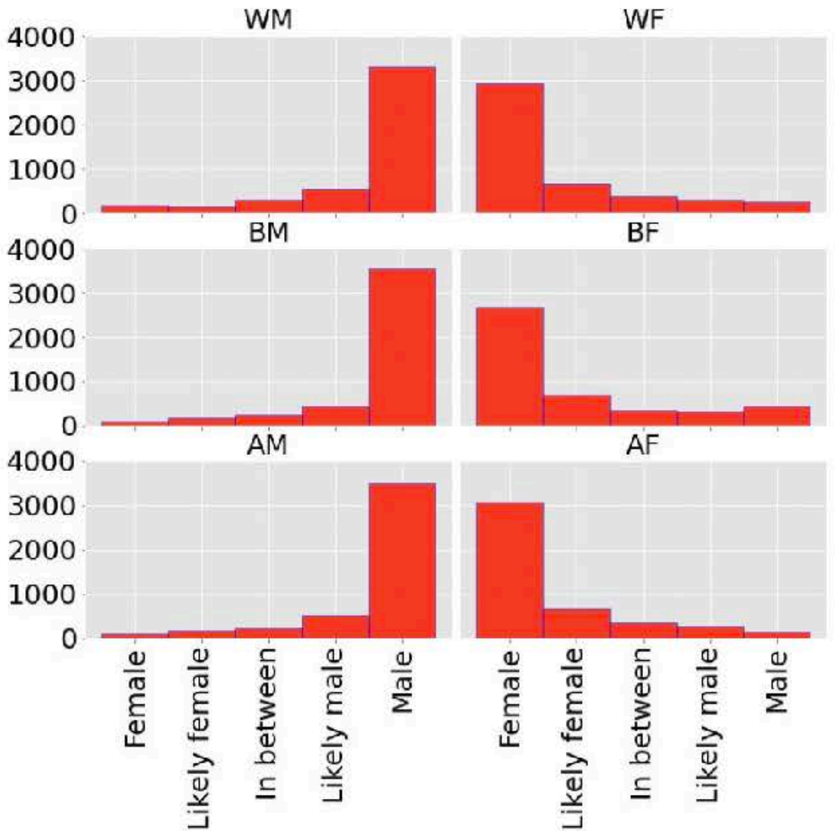
(a)

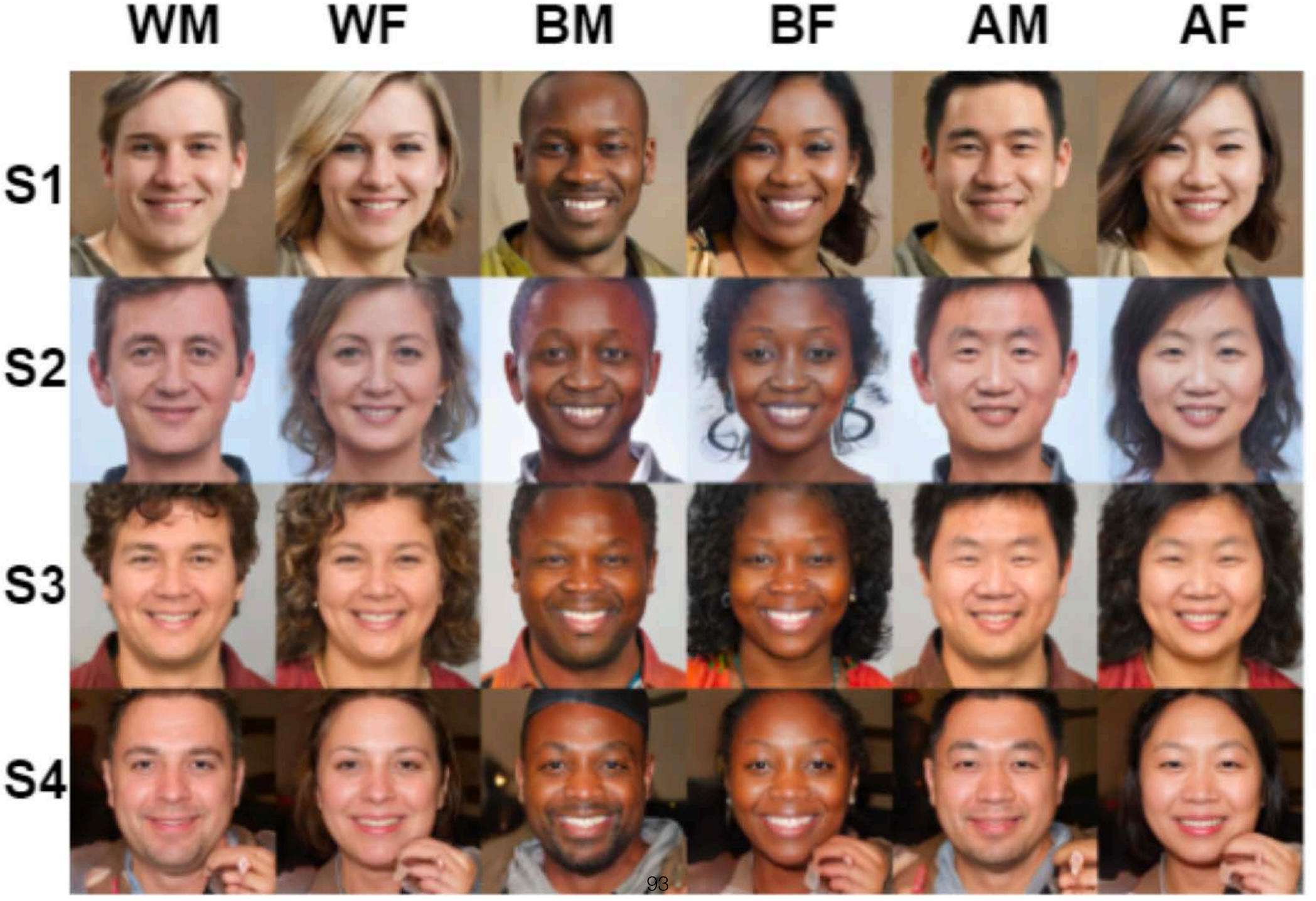

age group: 0
age group: 1
age group: 2
age group: 3
age group: 4
Vote count
4000
3000
2000
1000
0
Child
Teen
Young adult
Adult
Middle age
Senior

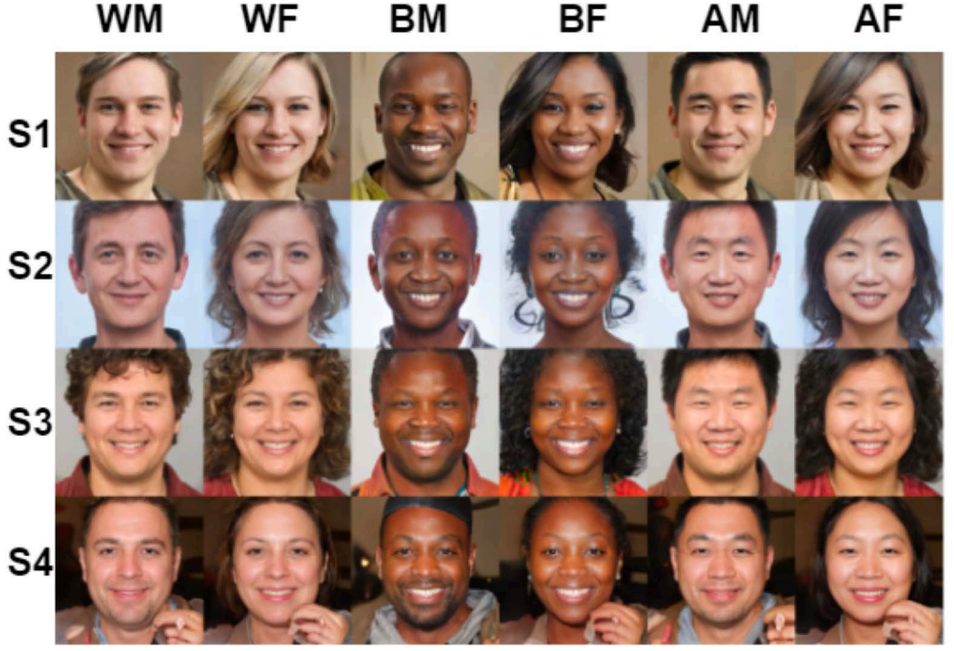
(b)

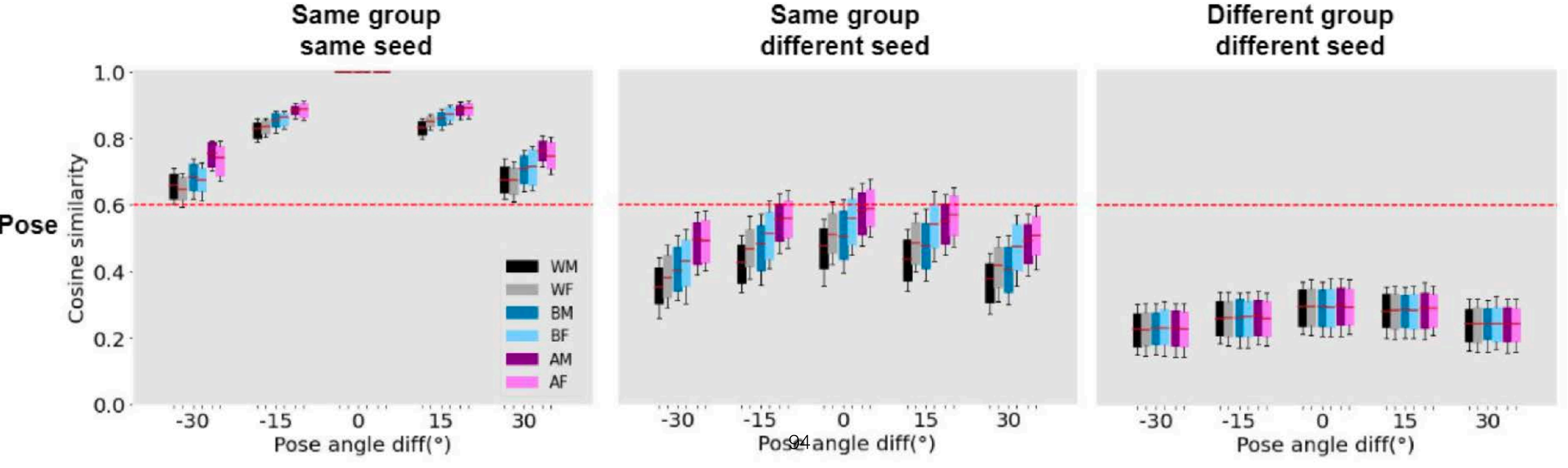

expression group: 0
expression group: 1
expression group: 2
expression group: 3
expression group: 4
2500
2000
1500
1000
500
0
Broad smile
Smile
Neutral
Serious
Frown

(c)


WM
WF
BM
BF
AM
AF
4000
3000
2000
1000
0
Female
Likely female
In between
Likely male
Male


WM
WF
BM
BF
AM
AF
S1
S2
S3
S4


WM
WF
BM
BF
AM
AF
S1
S2
S3
S4


Same group
same seed
Same group
different seed
Different group
different seed
Pose
Cosine similarity
1.0
0.8
0.6
0.4
0.2
0.0
WM
WF
BM
BF
AM
AF
-30
-15
0
15
30
Pose angle diff(°)

Instructions | Shortcuts | Same or different person? Are the two people the same (OK different age, hair style, facial expression), or different (maybe a close relative dressed and sty

Instructions

Likely same - Most people would think that they are the same.

Possibly same - Many but not all people would think that they are the same.

Not sure - People would be split on this.

Possibly different - Perhaps siblings. Many but not all people would think different.

Likely different - Most people would say different.

Select an option

Likely same	1
Possibly same	2
Not sure	3
Possibly different	4
Likely different	5

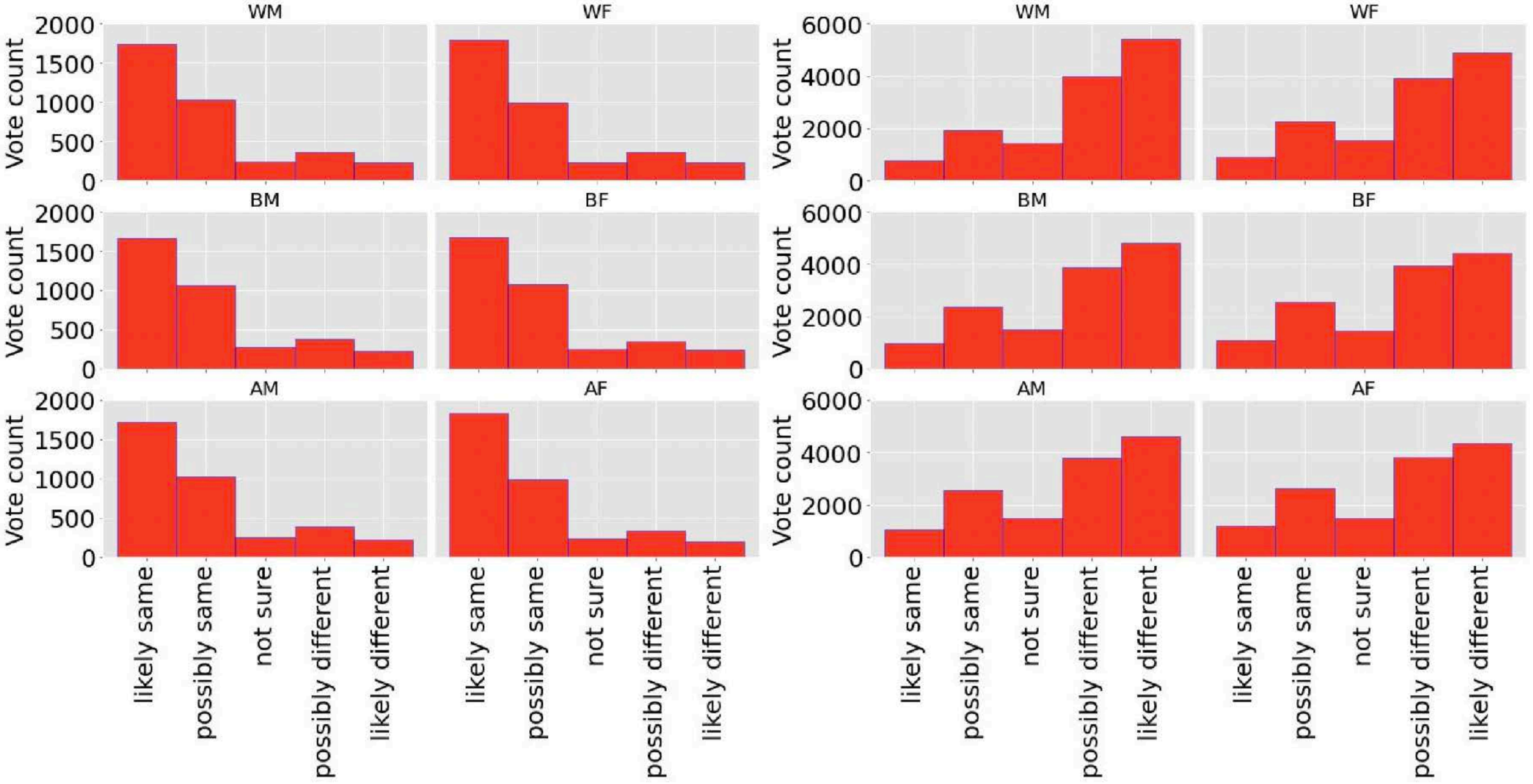
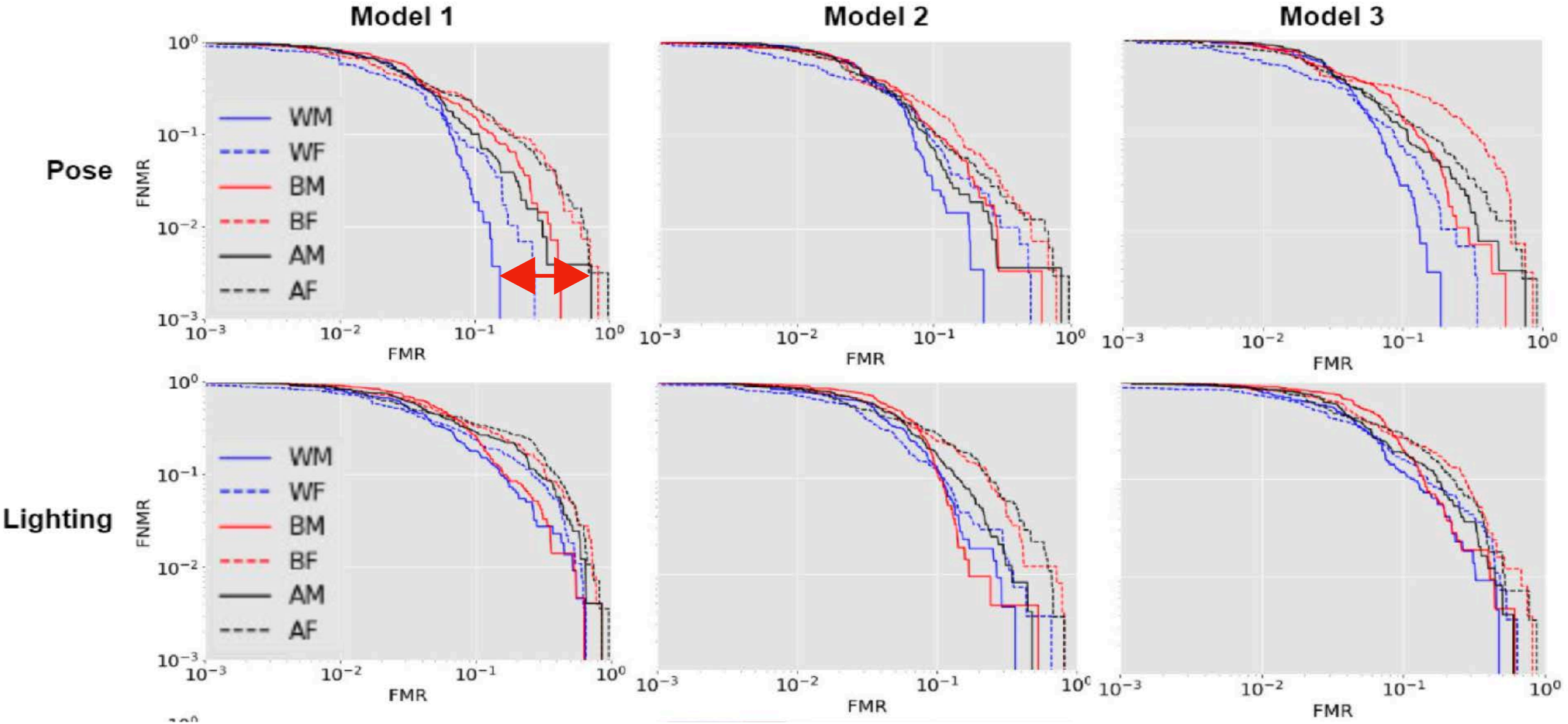


Figure 4. **Human annotation results on different pose image pairs. (a):** face images of different poses from the same seed and same demographic group. **(b):** face images of different poses from different seed but same demographic group. This trend is consistent across attributes, see Supplementary for results on other attributes.


Model 1
Model 2
Model 3
Pose
Lighting
WM
WF
BM
BF
AM
AF
FNMR
FMR
10^{0}
10^{-1}
10^{-2}
10^{-3}

Towards *responsible* responsible AI

- Observational methods 
- Experimental method 
- Synthetic samples + human annotations

Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale

Federico Bianchi*[1], Pratyusha Kalluri*[1], Esin Durmus*[1], Faisal Ladhak*[1,2], Myra Cheng*[1], Debora Nozza[3], Tatsunori Hashimoto[1], Dan Jurafsky†[1], James Zou†[1], and Aylin Caliskan†[4]

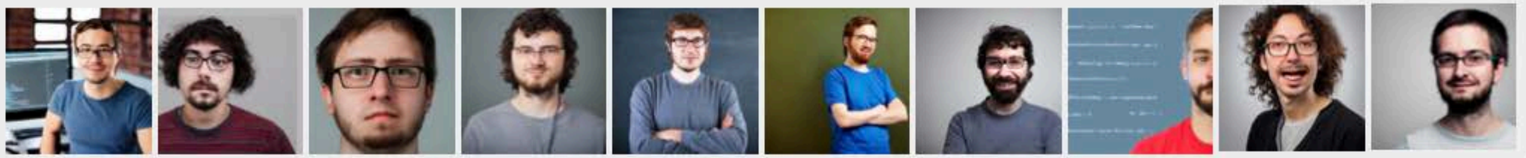
[1]Stanford University
[2]Columbia University
[3]Bocconi University
[4]University of Washington

Abstract

Machine learning models are now able to convert user-written text descriptions into naturalistic images. These models are available to anyone online and are being used to generate millions of images a day. We investigate these models and find that they amplify dangerous and complex stereotypes. Moreover, we find that the amplified stereotypes are difficult to predict and not easily mitigated by users or model owners. The extent to which these image-generation models perpetuate and amplify stereotypes and their mass deployment is cause for serious concern.

1 Introduction

a software developer



—> Young, pale, male, facial hair, eyeglasses

a flight attendant

—> Female, young, pale

a chef

—> Male

a cook

—> Female, brown

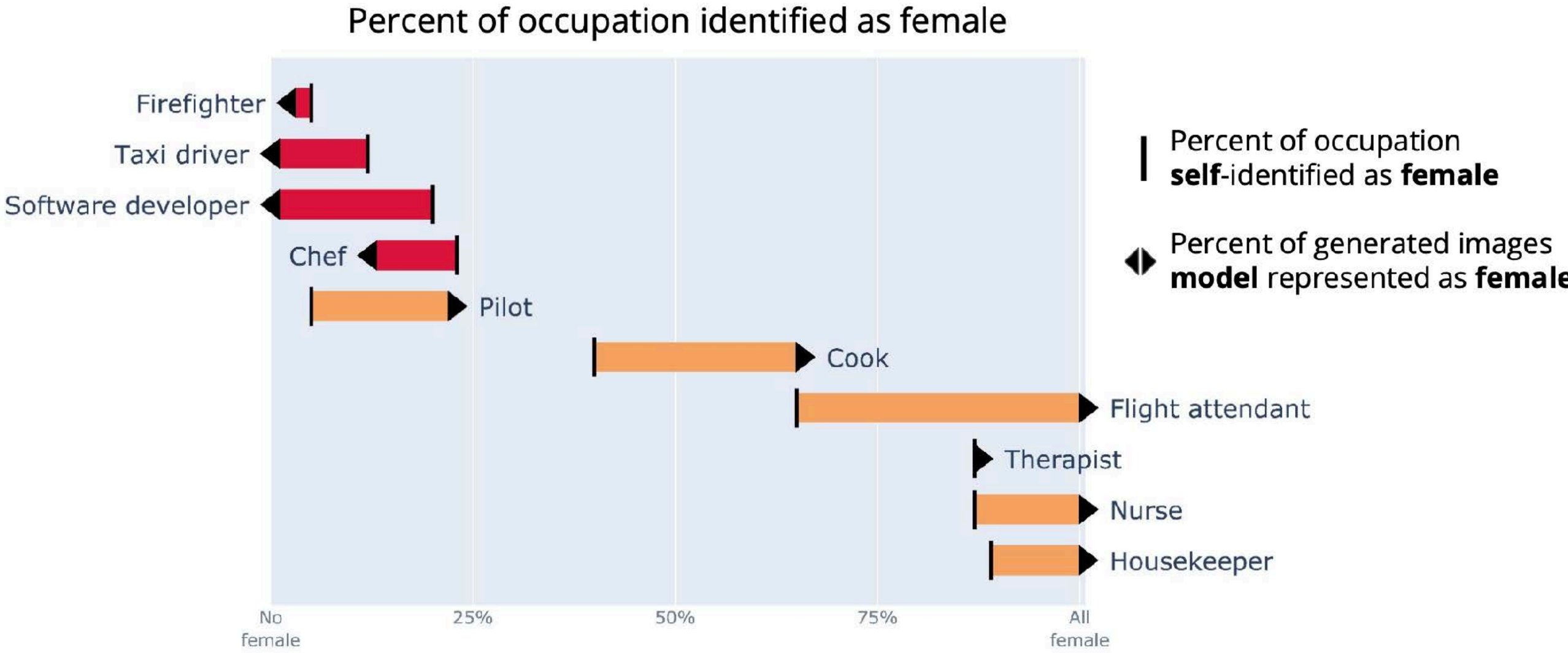
a taxi driver

—> Male, brown

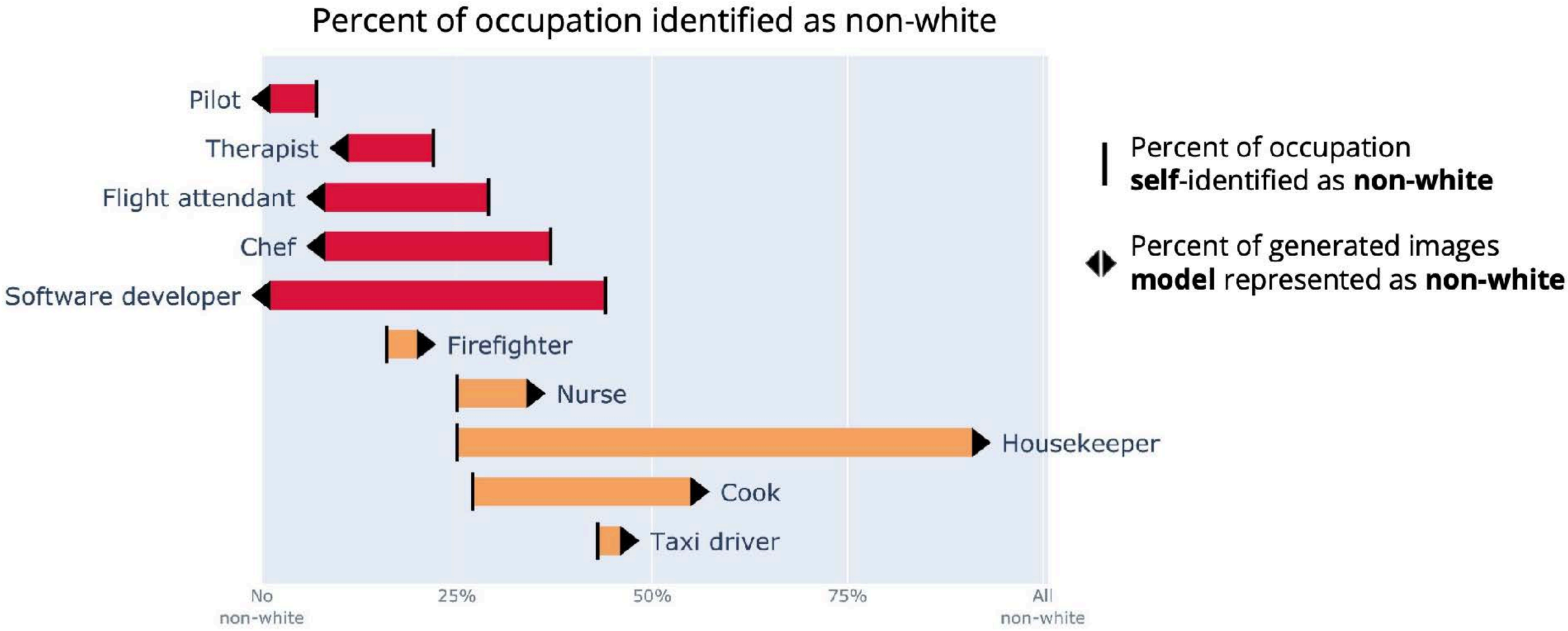
a housekeeper

—> Female, brown

Comparison with bureau of labor data



Comparison with bureau of labor data



Improvability

The New York Times

ECONOMIC VIEW

Biased Algorithms Are Easier to Fix Than Biased People

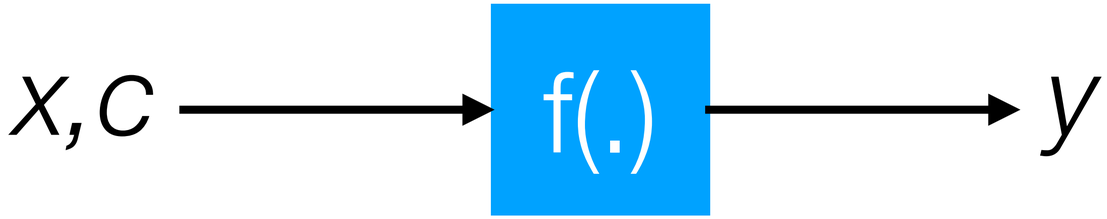
Racial discrimination by algorithms or by people is harmful — but that's where the similarities end.

By Sendhil Mullainathan

Dec. 6, 2019

Fixing bias?

Modify the loss function



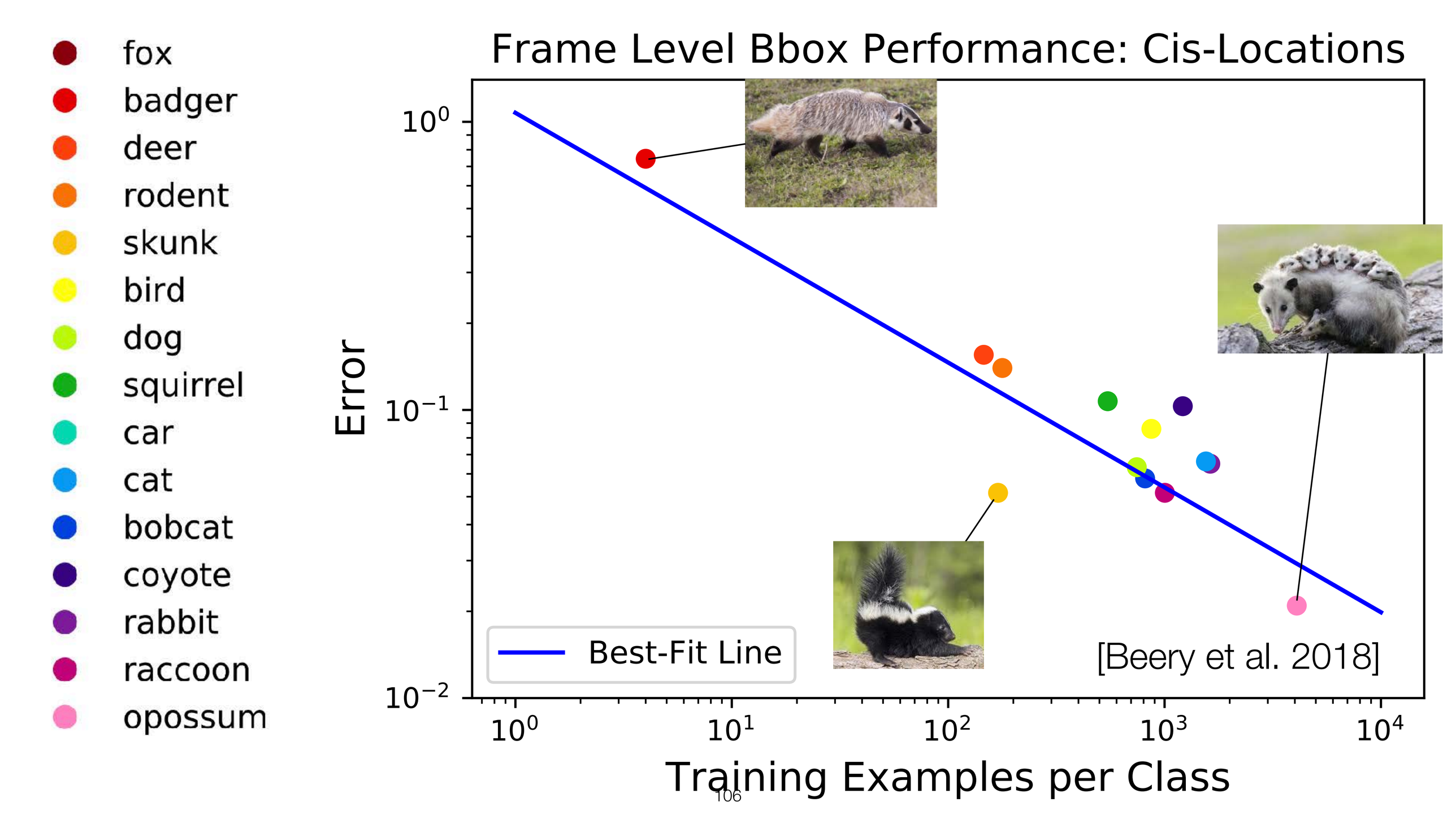
x = action, state

c = protected attribute

y = reward, outcome

$$p(y = 1|x, c) = p(y = 1|x)$$

$$p(y = 1|x, c = 0) = p(y = 1|x, c = 1)$$

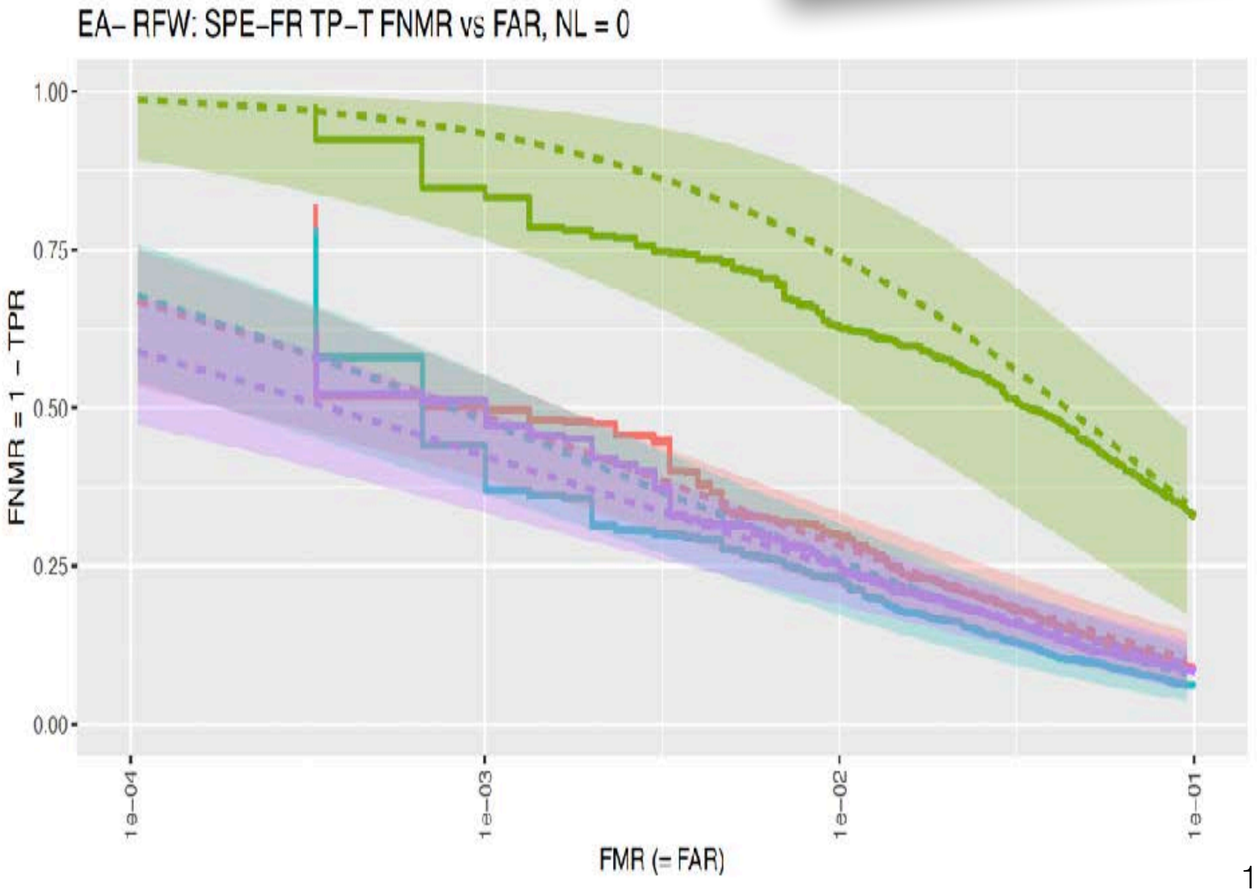

fox
badger
deer
rodent
skunk
bird
dog
squirrel
car
cat
bobcat
coyote
rabbit
raccoon
opossum
Frame Level Bbox Performance: Cis-Locations
Error
10^0
10^{-1}
10^{-2}
10^0
10^1
10^2
10^3
10^4
Training Examples per Class
Best-Fit Line
[Beery et al. 2018]

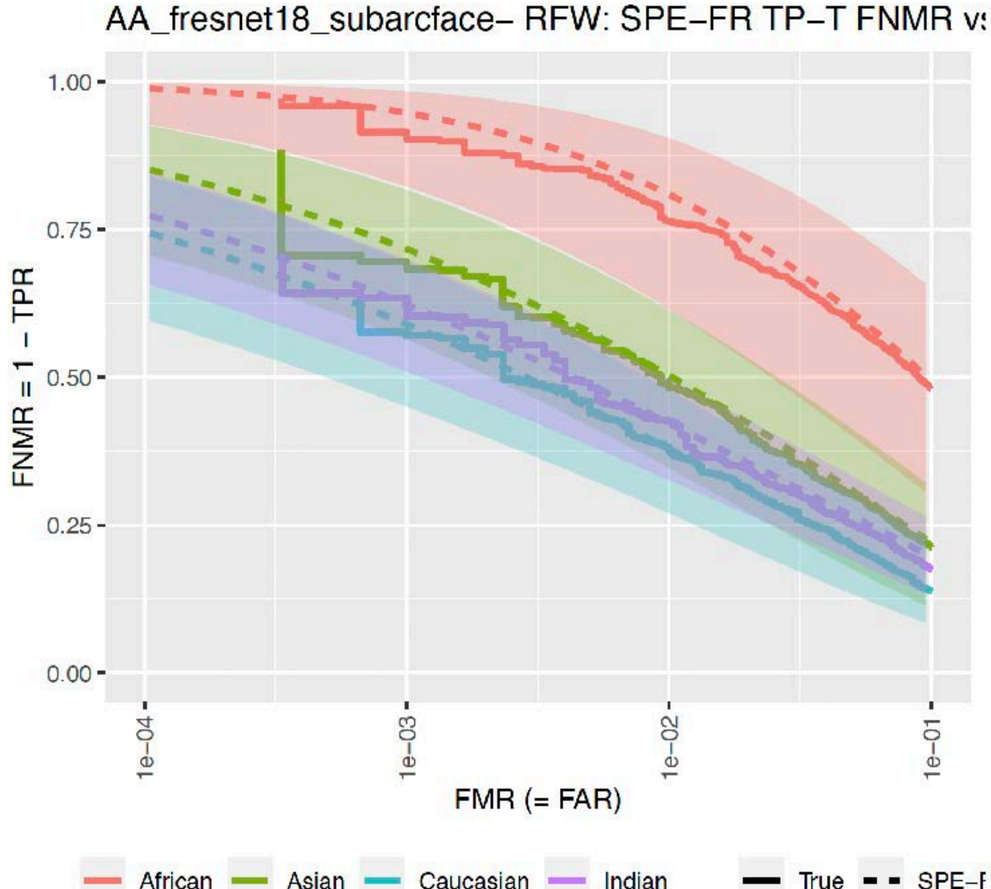
Unsupervised and Semi-supervised Bias Benchmarking in Face Recognition

Alexandra Chouldechova[†*] Siqi Deng[†] Yongxin Wang
Wei Xia[*] Pietro Perona

AWS AI Labs

ECCV2

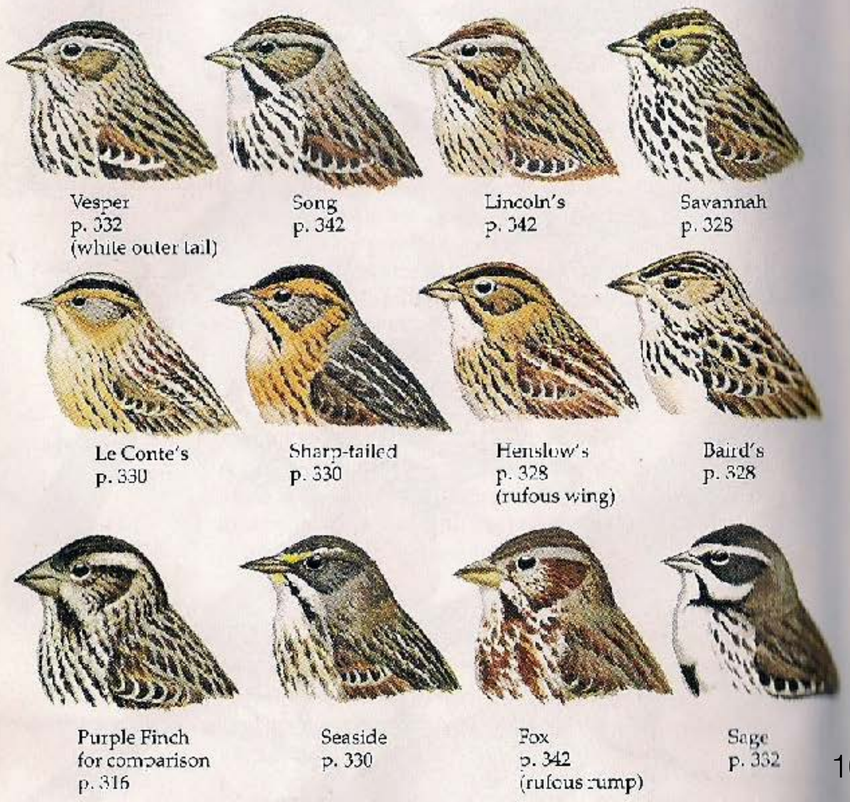




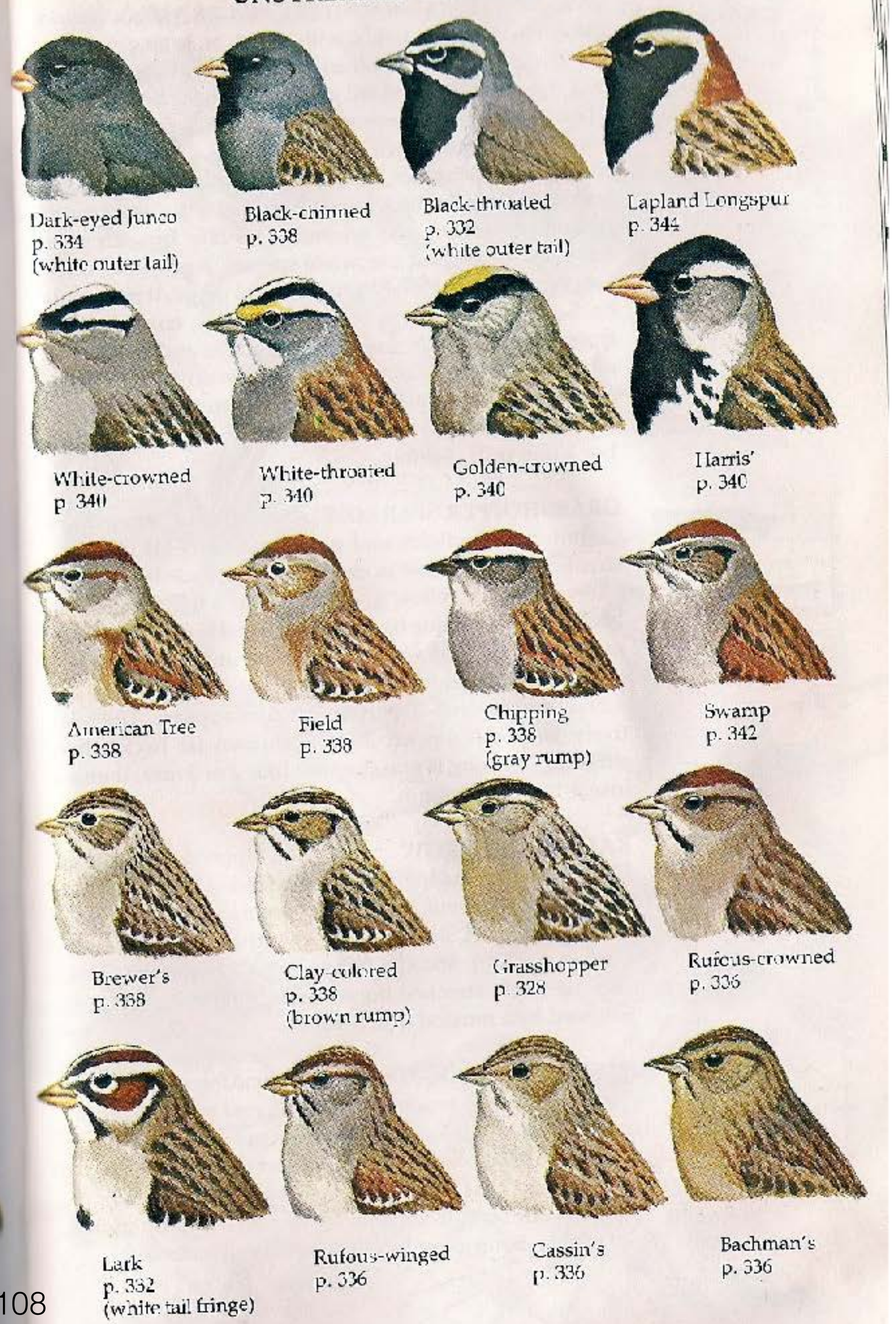
Sparrows

● **SPARROWS** are small brown-bodied birds with streaked backs and short conical beaks. Their food, mostly seeds except during the nesting season, is obtained on or near the ground. When not nesting, most are seen [illegible] these may be d[illegible] identifica- tion [illegible]es females are [illegible]pecies are sho[illegible]presented by one head each; the striking Lark Bunting, the Snow Bunting, the towhees, and the Olive Sparrow are omitted. Immatures of some species are much duller, especially those species with black or rufous on the head. Songs and chips of sparrows are often more easily distinguished than are their plumages. See pp. 328-345 for further details.

STREAKED BREASTS



UNSTREAKED BREASTS



Summary

- Responsible AI
- Bias and fairness
- Correlation vs causation
- Bias in face analysis and recognition
- Bias in generative models
- Fixing bias
- Conclusions

You are in the top 0.05% for training, and talent….

Your responsibilities

- Know, and follow, the law
- Aim for big positive outcomes (people, the planet)
- Be informed. Think about both good and bad outcomes of your work
- Engage with society: teach, debate, get involved