Analysis of Bias in AI Facial Beauty Regressors

Contents

Related Works

- Quer et al. (2024)
    - bias analysis framework demonstrated in hiring algorithms
        - Build model -> Test bias through analysis of predictions and errors
- Feldman and Peake (2021)
    - Formal definitions of Fairness adapted for regression
        - Distributional Parity
        - Error Parity
- Bias in Image Generation
    - less diverse outputs than human-curated content (Bogdanova et al., 2024)
    - Underrepresentation of women and People of Color in depictions of power and success (Gengler, 2024)
    - Whitifying non-white faces in image-to-image transformations (Yang, 2025)

Ethical implications

- Survey: Ilyas (2024)
    - 50% of respondents reported that exposure to AI-generated beauty standards negatively impacted their self-esteem
    - 70% agreed that such standards promote unrealistic cultural and social ideals
    - 82% of informants felt that AI-based beauty images are less inclusive in promoting diversity across cultures
- Philosophical Perspective: Zhou (2024)
    - Plato conceived of beauty as an abstract, eternal ideal while AI systems operationalize beauty in ways that are both highly specific and potentially exclusionary

Experiment

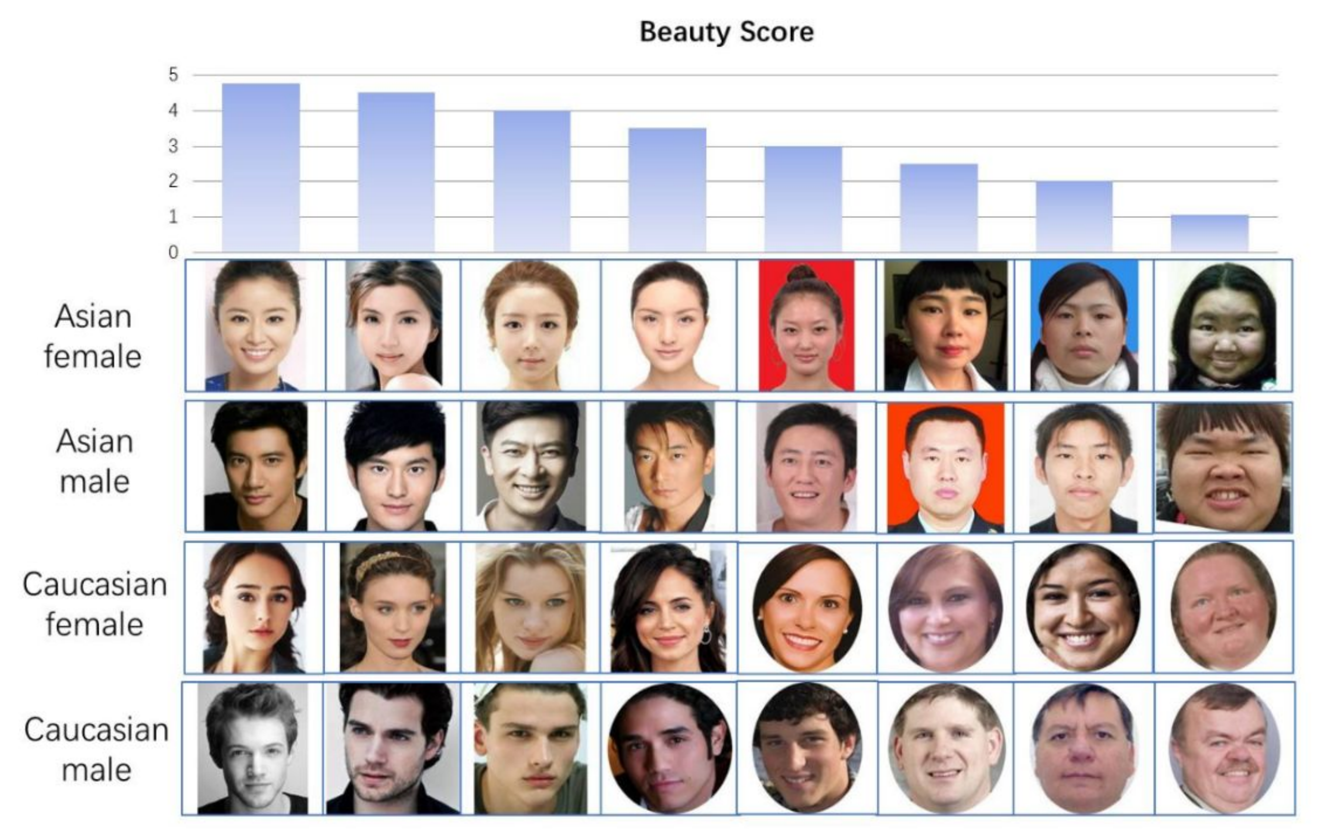
Three-phase computational pipeline

1. Model Development
    a. Fine-tune ResNet-152 architectures

2. Model Evaluation
    a. Generate beauty score predictions using each trained model

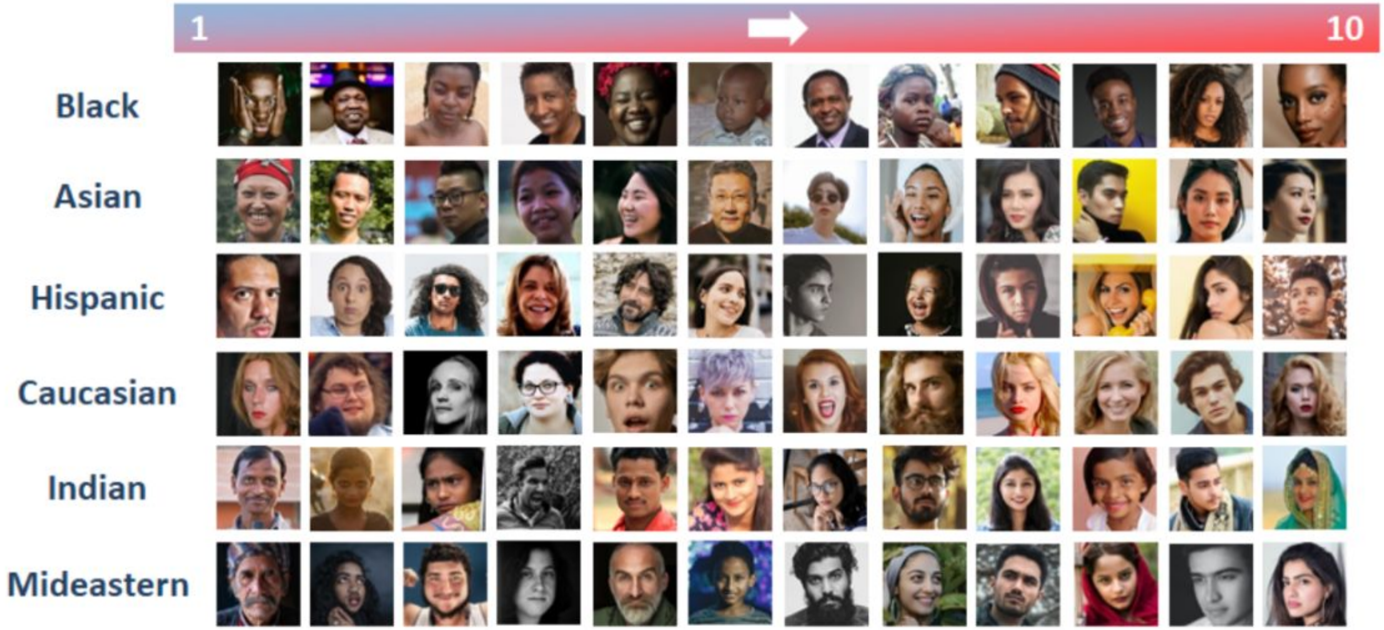
3. Bias Quantification
    a. Apply non-parametric statistical tests

Datasets (1 of 3)



SCUT-FBP5500
(Liang et al., 2018)

Datasets (2 of 3)



MEBeauty (Lebedeva et al., 2021)

Datasets (3 of 3)

FairFace Training Subset
(Karkkainen & Joo, 2021)

Preprocessing (Metadata)

	image	label	race	gender
0	CF437.jpg	0.500000	caucasian	female
1	AM1384.jpg	0.388393	asian	male
2	AM1234.jpg	0.303571	asian	male
3	AM1774.jpg	0.732143	asian	male
4	CF215.jpg	0.540178	caucasian	female
...	...	...	...	...
5495	AF546.jpg	0.232143	asian	female
5496	AM558.jpg	0.468750	asian	male
5497	AF805.jpg	0.383929	asian	female
5498	AF271.jpg	0.593750	asian	female
5499	AM1535.jpg	0.281250	asian	male

5500 rows × 4 columns

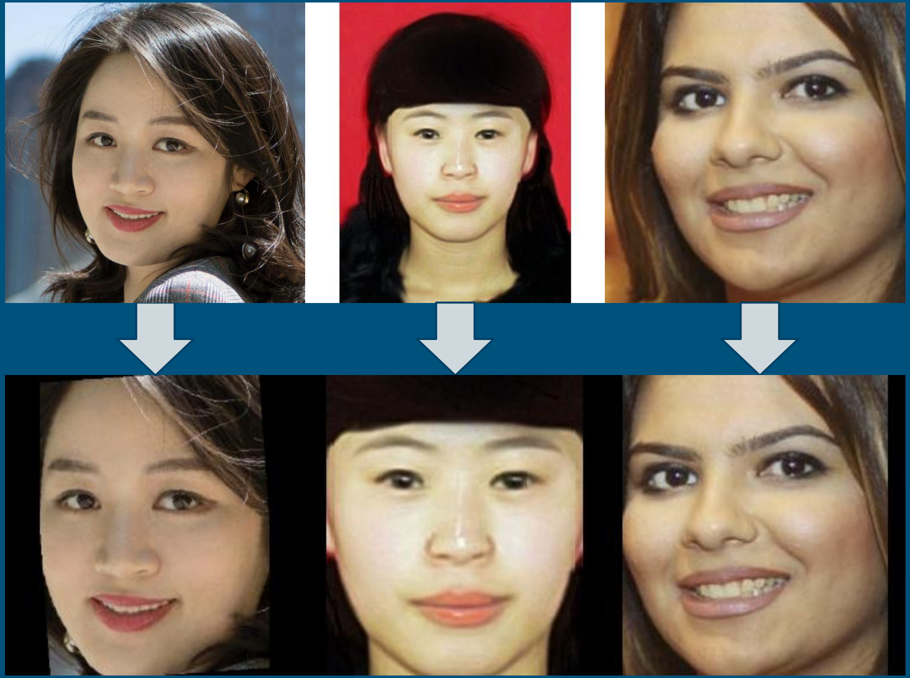
	label	image	gender	race
0	0.013640	kuma-kum-GKbPbR0ZAT4-unsplash.jpg	female	caucasian
1	0.000000	pexels-cottonbro-5529905.jpg	male	asian
4	0.057971	pexels-himesh-mehta-3059930.jpg	female	indian
5	0.103060	pexels-kaniseeyapose-2751061.jpg	male	asian
7	0.115942	imad-clicks-2_qmEnz7bQ4-unsplash.jpg	female	mideastern
...	...	...	...	...
2602	0.927536	pexels-pixabay-247322.jpg	female	caucasian
2603	0.971014	women-5930352_1920.jpg	female	asian
2604	0.953301	francesca-zama-1fhl_kmbfAE-unsplash.jpg	female	hispanic
2605	1.000000	sofia--LNdco1UgNY-unsplash.jpg	female	caucasian
2606	0.966184	pexels-mart¿¬-pardo-1674318.jpg	male	caucasian

2386 rows × 4 columns

	gender	race	image
0	male	east asian	1.jpg
1	female	indian	2.jpg
2	female	black	3.jpg
3	female	indian	4.jpg
4	female	indian	5.jpg
...	...	...	...
86738	male	middle eastern	86739.jpg
86739	male	indian	86740.jpg
86741	female	indian	86742.jpg
86742	female	black	86743.jpg
86743	male	white	86744.jpg

84729 rows × 3 columns

Preprocessing (Images)



MEB SCUT FairFace

Multi-task Cascaded Convolutional Network (MTCNN) face detector (Zhang et al., 2016)

Model Training (Data Loading)

- Augmentation: Random horizontal flips, ±10◦ rotations, and resized crops (80-100% scale) of training images
- Normalization: Pixel values scaled using ImageNet means (μ = [0.485, 0.456, 0.406]) and standard deviations (σ = [0.229, 0.224, 0.225])
- Datasets were split into training (2/3), validation (2/9), and test (1/9) subsets

Model Training (ResNet-152 Fine Tuning) (He et al., 2015)

Three Phases

1. Frozen Backbone
2. Unfreeze Conv5
3. Unfreeze Conv4

layer name	152-layer
conv1	
conv2_x	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times 3$
conv3_x	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times 8$
conv4_x	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix} \times 36$
conv5_x	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix} \times 3$

Bias Analysis (Model Performance)

Model	Test MSE	Cross-dataset MSE
SCUT-trained	0.008	0.024
MEBeauty-trained	0.013	0.028

Bias Analysis (MEBeauty Data)

	Predictions		Errors	
Race	Mean	Median	Mean	Median
Asian	0.56	0.56	-0.02	-0.03
Black	0.56	0.56	0.04	0.04
Caucasian	0.65	0.66	0.03	0.03
Hispanic	0.63	0.64	0.00	-0.00
Indian	0.60	0.61	-0.02	-0.03
Middle Eastern	0.66	0.66	0.07	0.07
KW Test Statistic	229		101	
KW Test P-value	$< 10^{-3}$		$< 10^{-3}$	

Bias Analysis (SCUT Data)

	Predictions		Errors	
Ethnic Group	Mean	Median	Mean	Median
Asian	0.5929	0.5938	0.0722	0.1004
Caucasian	0.5883	0.5859	0.0380	0.0582
MWU Statistic	2,869,019		2,562,921.5	
MWU P-value	0.012		$< 10^{-3}$	
KS Statistic	0.0702		0.146	
KS P-value	$< 10^{-3}$		$< 10^{-3}$	

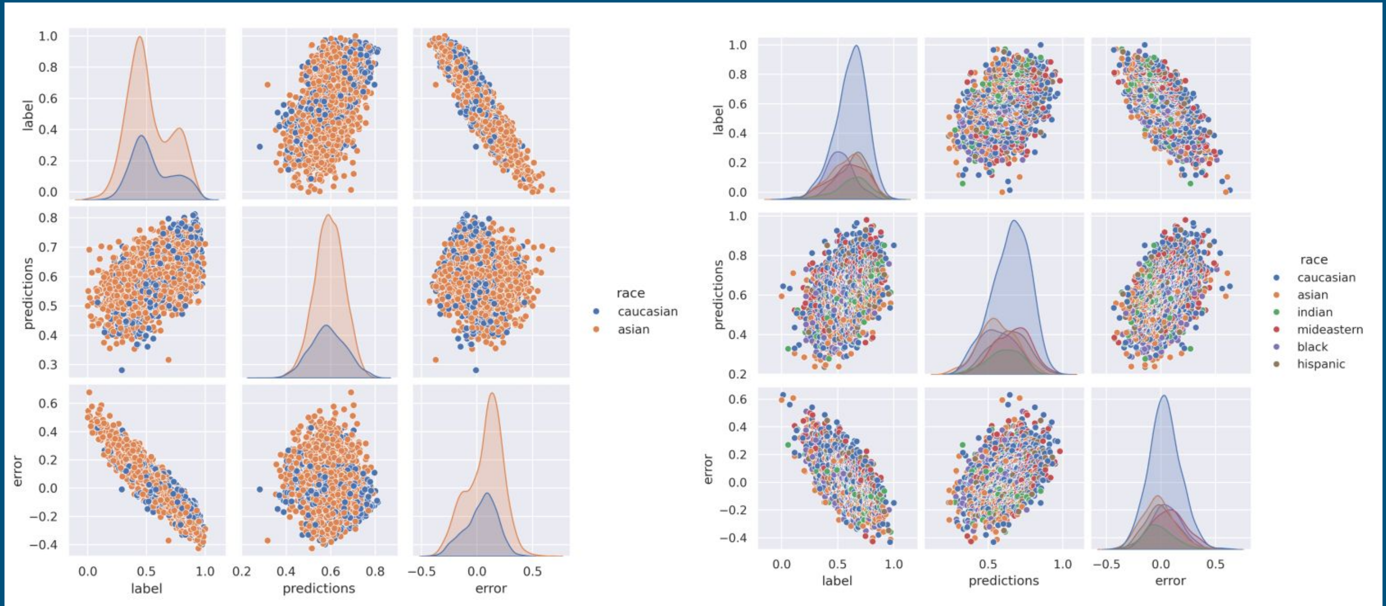
Bias Analysis (FairFace Data)

	SCUT Model		MEBeauty Model	
Ethnic Group	Mean	Median	Mean	Median
Black	0.460	0.461	0.525	0.527
East Asian	0.470	0.469	0.563	0.566
Indian	0.493	0.494	0.550	0.555
Hispanic	0.479	0.479	0.548	0.551
Middle Eastern	0.500	0.500	0.555	0.559
Southeast Asian	0.455	0.455	0.547	0.551
White	0.479	0.477	0.556	0.559
KW Test Statistic	1675.7		1716.8	
KW Test P-value	$< 10^{-3}$		$< 10^{-3}$	

https://github.com/Feebami/AI_Beauty_Bias/blob/main/BeautyBias_Hamel.pdf

Discussion (Sources of Bias)

- Sampling Bias
- Labeling Bias



Discussion (Ethical Implications and Paths Forward)

- Both models exhibited exacerbated bias on the balanced FairFace dataset
- The fairness criteria is stringent, but it should be the goal for responsible AI deployment
- Algorithmic Mitigation
    - Integration of fairness constraints during training
    - Yik and Silva (2024) and Yazdani-Jahromi et al. (2024)
- Data Curation
    - Stratified Sampling
    - Annotator diversity quotas
    - Metrics to capture cultural relativity
- Validation Protocols
    - Bias testing as part of model validation
    - Transparency

Conclusion

This study demonstrates that facial beauty prediction models have the potential to systematically encode ethnic biases

These biases stem from compounded representation and annotation limitations in beauty datasets

Unchecked deployment risks cementing algorithmic beauty standards that erase cultural diversity