Fairness at the Edge

Benchmarking Bias in Computer Vision Models

Dr Wiebke Hutiri, Senior Research Scientist at Sony Group

AI for Good Workshop on *Intelligence at the Edge*
10 July 2026

Bias in HCCV can lead to real-world harms

A Wrongful Arrest and Worry About the Accuracy of a Police Tool

Trevis Williams was arrested after facial recognition identified him as a possible match for a suspect. “In the blink of an eye, your whole life could change,” he said.

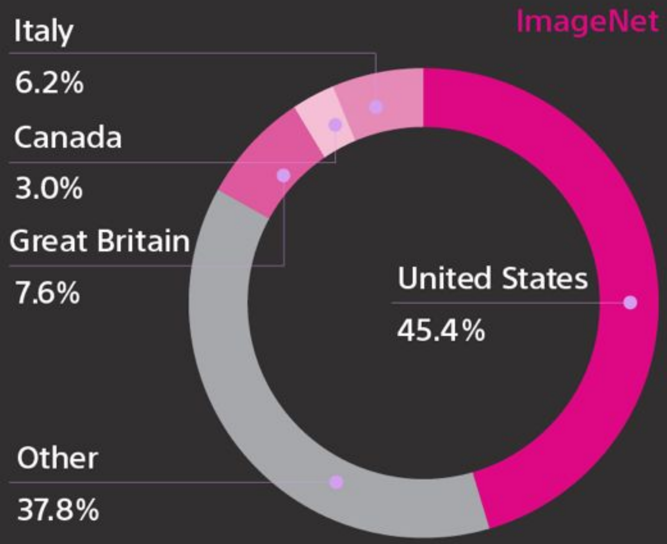
Driverless cars worse at detecting children & darker-skinned pedestrians say scientists

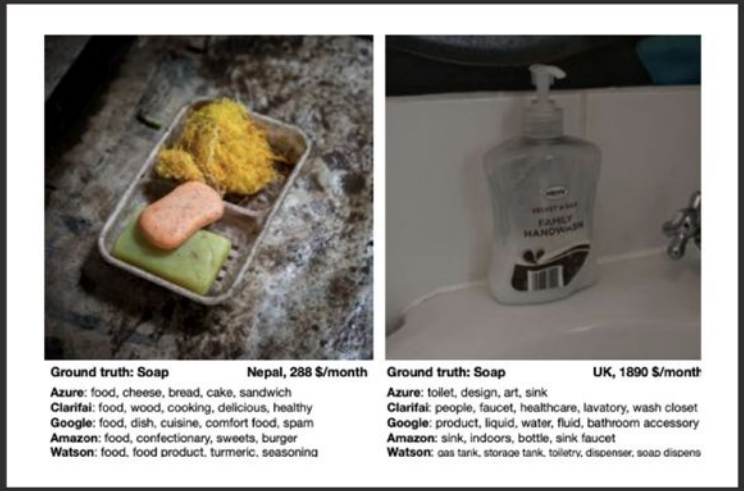
Researchers call for tighter regulations following major age and race-based discrepancies in AI autonomous systems.

Is the iPhone X Racist? Apple Refunds Device That Can’t Tell Chinese People Apart, Woman Claims

Faulty face verification systems can lead to security risks.

Data selection molds an AI system’s world view





Major academic image datasets are sourced primarily from a few developed countries [1]

Object recognition models perform poorly on images from lower-income countries [2]

AI systems suffer from “own-race” bias in recognition, similar to humans [3]

[1] Shankar et al. “No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World.” NeurIPS Workshop 2017.
[2] De Vries et al. “Does object recognition work for everyone?” CVPRW 2019.
[3] Phillips et al. “An Other-Race Effect for Face Recognition Algorithms.” ACM Transactions on Applied Perception 2011.s

Data curation has historically been under-valued

Bespoke image collection in studios led to controlled but small and constrained datasets [1].

ImageNet popularized web-scraping, revolutionizing the development of large-scale image datasets [2].

Generative AI and AI agents further raises importance of developing ethical data collection practices.

[1] Raji and Fried. “About Face: A Survey of Facial Recognition Evaluation.” arXiv 2021.
[2] Deng et al. “Imagenet: A Large-Scale Hierarchical Image Database.” CVPR 2009.

FHIBE:

Fair Human-centric Image BEnchmark

The first publicly available, globally diverse, consensually-collected fairness evaluation dataset for a wide variety of human-centric computer vision tasks

Alice Xiang, Jerone T. A. Andrews, Rebecca L. Bourke, William Thong, Julienne M. LaChance, Tiffany Georgievski, Apostolos M
Rahmattalabbi, Yunhao Ba, Shruti Nagpal, Orestis Papakyriakopoulos, Dora Zhao, Jinru Xue, Victoria Matthews, Linxia Gong, A
Mircea Cimpoi, Swami Sankaranarayanan, Wiebke Hutiri, Morgan K. Scheuerman, Albert S. Abedi, Peter Stone, Peter R. Wurr
Kitano & Michael Spranger.

Fair human-centric image dataset for ethical AI benchmarking. Nature (2025).
https://doi.org/10.1038/s41586-025-09716-2

Made for Ethics and Built with Ethics

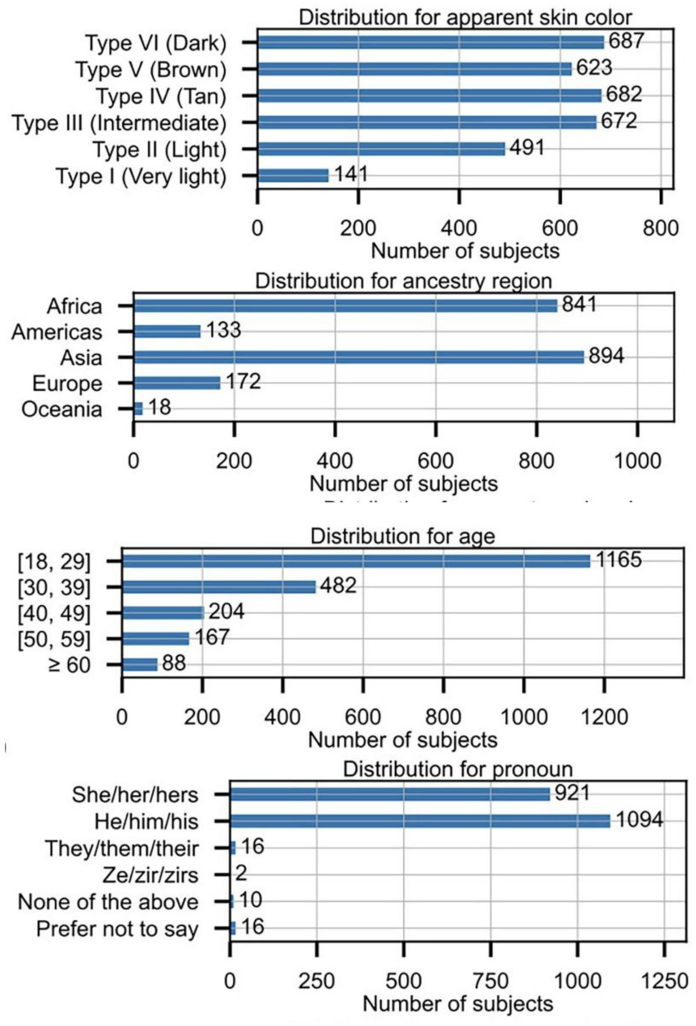
Made for AI ethics evaluation

- A practical dataset that enables AI developers to **diagnose bias** in AI models
- Can be applied to a wide range of **human-centric computer vision tasks**, including person detection, pose estimation, and face parsing.
- **Richer annotations** than any existing publicly available computer vision dataset, meaning **more granular and accurate** bias diagnoses.

Built in an ethically robust way

- A dataset that is **ethically sourced and curated**, and doesn't rely on problematic data
- First public **GDPR-compliant** human-centric computer vision dataset.
- Emphasis on **informed consent and fair compensation** for data subjects and annotators. Consent can be withdrawn at any time.

Made for Ethics and Built with Ethics



- 785 camera models
- 49 camera manufacturers
- 16 scene types
- 6 lighting conditions
- 7 weather scenarios
- 3 camera positions
- 5 camera distances
- 16 body poses
- 2 head poses
- 47 distinct subject-subject and subject-object interactions
- 15 hair and 4 facial hair styles
- 7 hair types
- 13 hair and 12 facial hair colors
- 9 eye colors
- 11 types of facial marks
- 6 pronoun categories
- 56 integer ages (18 to 75 years)
- 22 ancestry subregions within 5 regions
- 6 Fitzpatrick skin tones
- 33 keypoints
- 28 segmentation categories

Principles Underlying the Design of FHIBE

Consent

Privacy & IP

Fair Compensation

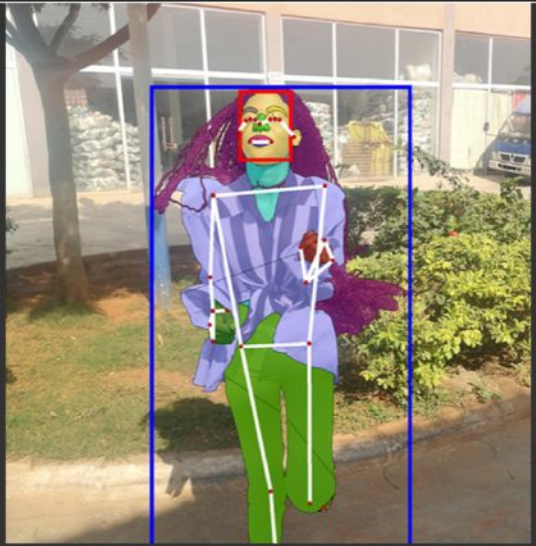
Safety

Diversity

FHIBE comprises 10,319 images of 1,981 unique subjects from 81 countries/areas.

This includes:

Image subject annotations, including self-reported annotations
(e.g. age, ancestry)



This includes:

Instrument annotations
(e.g. capture time, camera shutter speed)

This includes:

Environment annotations
(e.g. scene, lighting)

Plus: Pixel-level annotations

(e.g. face and person bounding boxes, key points, and segmentation masks)

FHIBE comprises 10,319 images of 1,981 unique subjects from 81 countries/areas.

Images feature both ***primary*** and ***secondary* subjects**, increasing the diversity and complexity of images

1,234 intersectional groups (defined by age group, pronoun, ancestry subregion, and Fitzpatrick skin tone)

Plus: FHIBE includes **two derivative face datasets**: (1) a cropped-only set, and (2) a cropped-and-aligned set.

Comparing FHIBE to Existing Datasets

Extended Data Table 1 | Overview of human-centric computer vision (HCCV) datasets commonly used for fairness

Dataset	Size	Collection Method	Collection Detail	BB	KP	SM	Consent	Terms of Use	Demographic Diversity
MS-Celeb-1M	10M	scraped	web	-	-	-	-	revoked	**Nationality:** >200 **Race:** 3 groups **Gender:** binary
YFCC100M	99.2M	scraped	Flickr	-	-	-	-	revoked	**Geo:** coordinates
Megaface	4.7M	derived	YFCC100M	aF	a^{3}	-	-	revoked	**Geo:** coordinates
VGGFace	3.31M	scraped	Google Images	-	-	-	-	revoked	**Gender:** binary
Diversity in Faces (DiF)	1M	derived	YFCC100M	aF	a^{19}	-	-	revoked	**Skin color:** continuous **Gender:** binary **Age:** 7 groups
Pilot Parl. Benchmark	1.27k	sampling	government websites	-	-	-	-	revoked	**Skin tone:** 6 groups **Gender:** binary
FRGC	50K	direct	university lab	-	-	-	-	revoked	**Race:** 3 groups **Sex:** binary **Age:** 3 groups
RFW	41k	derived (scraped)	MS-Celeb-1M	-	-	-	-	n-c research	**Race:** 4 groups **Gender:** binary
Morph	55k	sampling	public records	-	-	-	-	n-c	**Race:** 4 groups **Gender:** binary **Age:** 4 groups
Adience	26.6k	scraped	Flickr	-	-	-	-	research	**Gender:** binary **Age:** 8 groups
BUPT-Globalface	2M	derived (scr.), scraped	MS-Celeb-1M, Google Images	-	-	-	-	n-c research	**Race:** 4 groups **Gender:** binary
WIDERFACE-DEMO	16k	derived (scraped)	WIDER FACE	mF	-	-	-	CC BY-NC 4.0	**Skin tone:** 6 groups **Gender:** 3 groups **Age:** 6 groups
KANFace	41k	scraped	YouTube	aF	-	-	-	n-c research	**Gender:** binary **Age:** discrete
FairFace	108k	derived (scraped)	YFCC100M	aF	-	-	-	CC BY 4.0	**Race:** 7 groups **Gender:** binary **Age:** deciles
ImageNet (ILSVRC)	1.4M	scraped	search engines	mO	-	-	-	edu., n-c research	-
CelebA	202.6k	derived (scraped)	CelebFaces	-	m^{5}	-	-	n-c research	**Skin tone:** binary **Gender:** binary **Age:** binary
LFWA	13.2k	derived (scraped)	LFW	-	m5	-	-	-	**Skin tone:** binary **Gender:** binary **Age:** binary
MTFL	13k	derived (scraped), scraped	LFW, web	-	m^{5}	-	-	-	**Gender:** binary
UTKFace	20k	derived, scraped	Morph, CACD	-	a^{68}	-	-	n-c research	**Race:** 5 groups **Gender:** binary **Age:** discrete
MIAP	100k	derived (scraped)	Open Images V6	mP	-	v^{350}	-	-	**Gender:** 3 groups **Age:** 3 groups
FACET	32k	derived (licensed)	SA-1B	mP	-	a^{3}	-	n-c research eval.	**Skin tone:** 11 groups **Gender:** 4 groups **Age:** 4 groups
MS-COCO	328k	scraped	Flickr	aP	m^{18}	m^{91}	-	-	-
VQA 2.0	200k	derived (scr.)	MS-COCO	aP	m^{18}	m^{91}	-	-	-
Casual Conversations	45k videos	crowd-sourced	vendors	-	-	-	-	eval.	**Skin tone:** 6 groups ***Gender:** 3 groups ***Age:** 3 groups
CCV2	26.5k videos	crowd-sourced	vendors	-	-	-	no details	eval.	**Skin tone:** 6, 10 groups ***Gender:** ***Age:** discrete **Geo.:** 7 countries
Chicago Face Database	158	direct	university lab	-	m^{189}	-	no details	n-c research	**Skin color:** continuous **Race:** 2 groups **Gender:** binary
Dollar Street	38k	direct	photographers	-	-	-	details	CC BY 4.0	**Geo.:** 63 countries
FHIBE (Our Dataset)	10k	crowd-sourced	vendors	mFP	m^{33}	m^{28}	details, for AI	eval.	***Skin tone:** 6 groups ***Ancestry:** 22 regions ***Pronouns:** 6 groups ***Age:** discrete ***Geo.:** 81 countries

The majority of prior HCCV datasets were **scraped** from Internet platforms, or derived from scraped datasets.

Six were revoked by their authors and are no longer publicly available: likely due to ethical or legal concerns.

Many lack dense pixel-level annotations, and those that have them contain limited demographic annotations.

Only four other HCCV datasets obtain consent: each one is either small, has limited or problematic consent strategies, or lacks pixel-level annotations, thereby limiting their utility.

FHIBE stands out as the only dataset collected with robust consent for AI evaluation and bias mitigation.

Comparing FHIBE to Existing Datasets

Table 1 | Dataset comparison by intersectional subgroups

Intersectional group	Dataset	Number of subgroups	Number of images		
			Med.	Max.	Min.
Gender × age	FHIBE	23	23	3,353	1
	CCv1	12	220	523	1
	CCv2	23	23	1,598	1
	FACET	9	2,070	22,008	3
	MIAP	4	7,439	21,195	254
Gender × age × skin tone	FHIBE	92	42	1,168	1
	CCv1	62	38	129	1
	CCv2	137	5	909	1
	FACET	82	284	12,506	1
Gender × age × ancestry region	FHIBE	72	128	8,415	4
Gender × age × ancestry subregion	FHIBE	322	31	1,683	1
Gender × age × ancestry region × skin tone	FHIBE	275	36	5,645	4
Gender × age × ancestry subregion × skin tone	FHIBE	1,234	9	1,129	1

Table 2 | Dataset comparison by summary statistics

	FHIBE	CCv1	CCv2	FACET	MIAP
Images/video frames	10,318	45,186	26,467	31,702	13,762
Subjects	1,981	3,011	5,567	NA	NA
Attributes	71	7	21	11	9
Attribute values	8,571	294	506	97	222

FHIBE stands out from other datasets in terms of its **detailed** and **self-reported demographic labels**.

Comparing FHIBE to Existing Datasets

Table 3 | Dataset comparison by geographical region and income level

<table>
<tr><th></th><th>FHIBE (%)</th><th>CCv1 (%)</th><th>CCv2 (%)</th><th>FACET (%)</th><th>COCO (%)</th><th>MIAP (%)</th></tr>
<tr><td>Africa</td><td>44.7</td><td>0.0</td><td>0.0</td><td>2.8</td><td>3.0</td><td>1.7</td></tr>
<tr><td>Asia and Oceania</td><td>40.6</td><td>0.0</td><td>49.8</td><td>36.2</td><td>11.4</td><td>14.3</td></tr>
<tr><td>Europe</td><td>4.4</td><td>0.0</td><td>0.0</td><td>49.8</td><td>34.2</td><td>36.2</td></tr>
<tr><td>Latin America and Caribbean</td><td>4.2</td><td>0.0</td><td>42.5</td><td>3.5</td><td>3.1</td><td>5.0</td></tr>
<tr><td>North America</td><td>6.0</td><td>100.0</td><td>7.7</td><td>7.7</td><td>48.3</td><td>42.8</td></tr>
<tr><td>High-income economies</td><td>11.5</td><td>100.0</td><td>7.7</td><td>54.0</td><td>89.1</td><td>87.5</td></tr>
<tr><td>Upper-middle-income economies</td><td>14.5</td><td>0.0</td><td>50.5</td><td rowspan="2">45.0</td><td rowspan="2">10.5</td><td rowspan="2">12.0</td></tr>
<tr><td>Lower-middle-income economies</td><td>71.5</td><td>0.0</td><td>41.8</td></tr>
<tr><td>Low-income economies</td><td>0.3</td><td>0.0</td><td>0.0</td><td>0.9</td><td>0.4</td><td>0.5</td></tr>
</table>

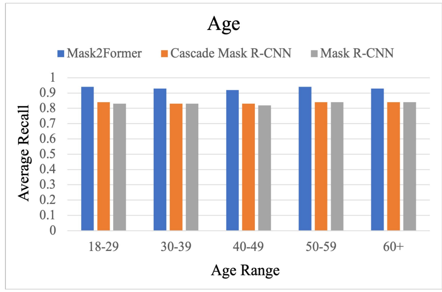
FHIBE has greater representation from regions that are underrepresented in many computer vision datasets, such as Africa (44.7%) and lower-middle income economies (71.5%).

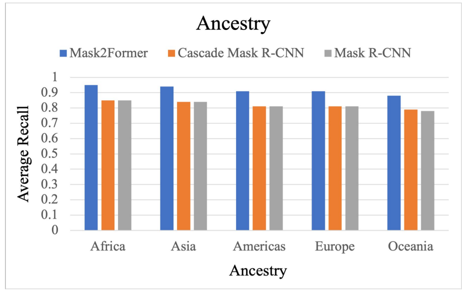
Improving Edge AI Robustness with Bias Evals

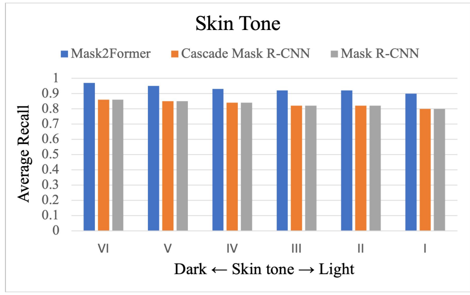
Benchmarking across 8 tasks (4 human-centric and 4 face-centric) across multiple datasets and models:

- Pose Estimation
- Person Segmentation
- Person Detection
- Face Detection
- Face Parsing
- Face Verification
- Face Reconstruction
- Face Super-resolution

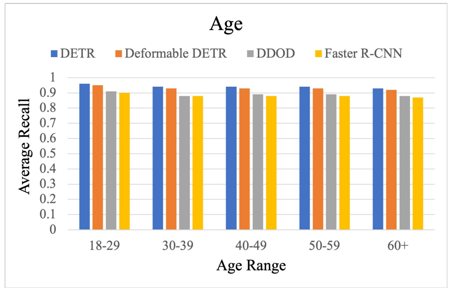
Person Segmentation

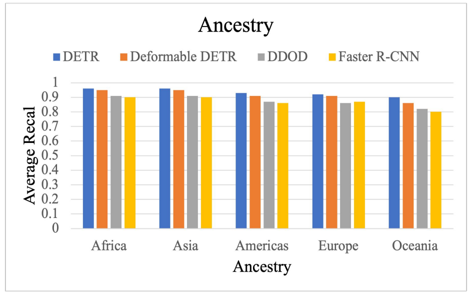


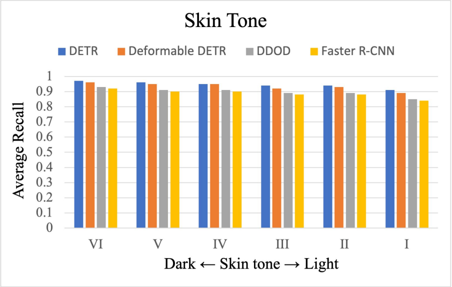




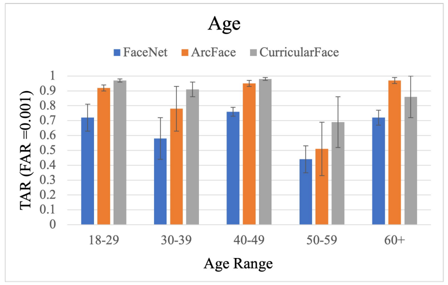
Person Detection

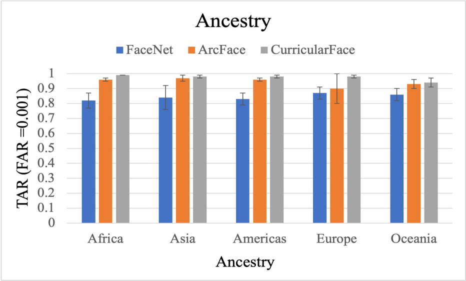


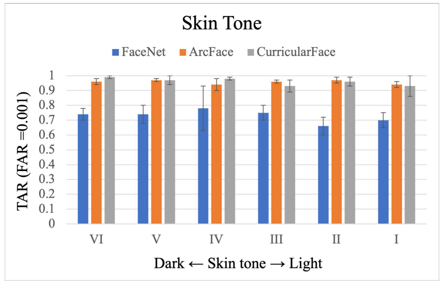




Face Verification

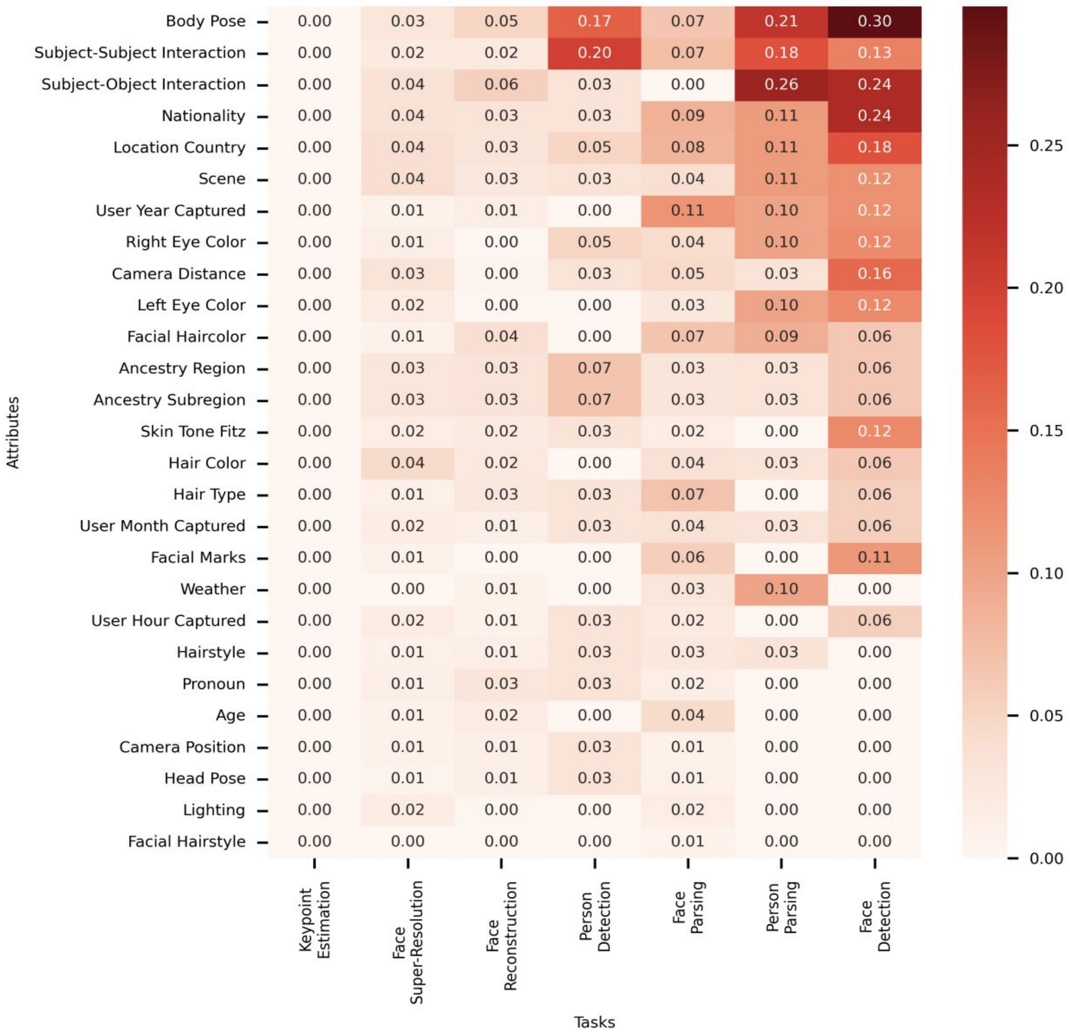






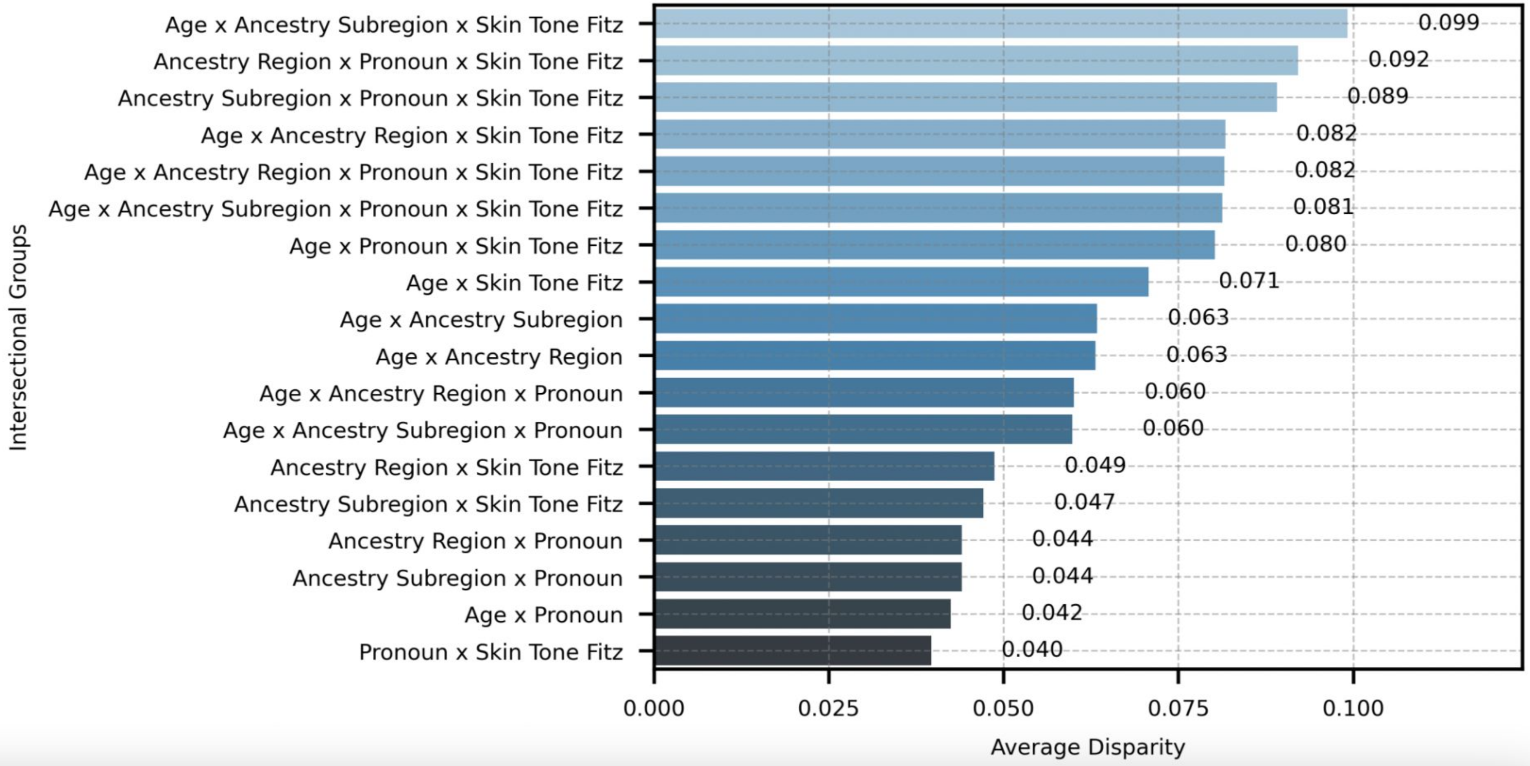
Improving Edge AI Robustness with Bias Evals

Model disparities can be assessed over
a vast range of attributes



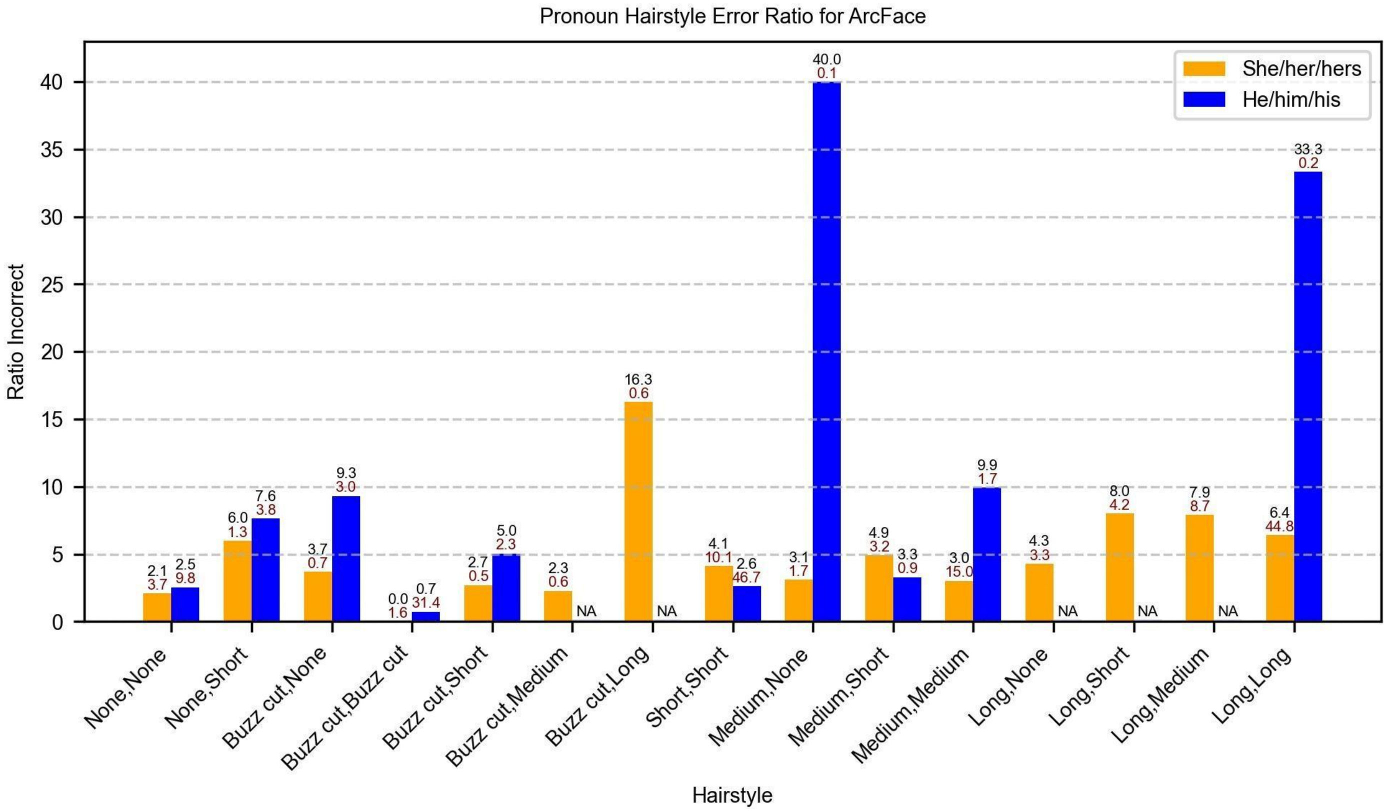
Improving Edge AI Robustness with Bias Evals

Intersectional groups combining multiple sensitive attributes experience the largest performance disparities



Improving Edge AI Robustness with Bias Evals

Difference in hairstyle leads to performance drop. He/Him/His has higher errors in not-stereotypical hairstyle (Long/None).



The number on top of each bar in black denotes the ratio of incorrect samples within that subgroup, while the number in red denotes the percentage of individuals with that pronoun who exhibit the corresponding hairstyle combination.

Download and use FHIBE:

http://fairnessbenchmark.ai.sony