Development of Trustworthy Image Classification Systems within a Sociotechnical Context

UNIVERSITY OF MIAMI

Rahul Dass

Ph.D. Student, U-LINK Predoctoral Fellow

August 22, 2022

Department of Computer Science, University of Miami

Thank you to my dissertation committee members

Dr. Ubbo Visser
(CS, Chair)

Dr. Nick Petersen
(Sociology & Law,
co-advisor)

Dr. Odelia Schwartz
(CS)

Dr. Victor Milenkovic
(CS)

Additional collaborators

Dr. Marisa Omori
(Criminology &
Criminal Justice, UMSL)

Dr. Tamara R. Lave
(Law)

Funding

U-LINK

UNIVERSITY OF MIAMI

Agenda

1. Trustworthy ML
    a. Why do we care?

2. Problem: unequal racial treatment
    a. Social justice issues
    b. Vision and ML-based FPT issues

3. Equitable DL methodology
    a. Data and tackling biases
    b. Phase 1: multidimensionality of race
    c. Phase 2: “self-auditing” evaluation

4. Discussion & Recommendations

UNIVERSITY OF MIAMI

Agenda

1. Trustworthy ML
   a. Why do we care?

2. Problem: unequal racial treatment
   a. Social justice issues
   b. Vision and ML-based FPT issues

3. Equitable DL methodology
   a. Data and tackling biases
   b. Phase 1: multidimensionality of race
   c. Phase 2: “self-auditing” evaluation

4. Discussion & Recommendations

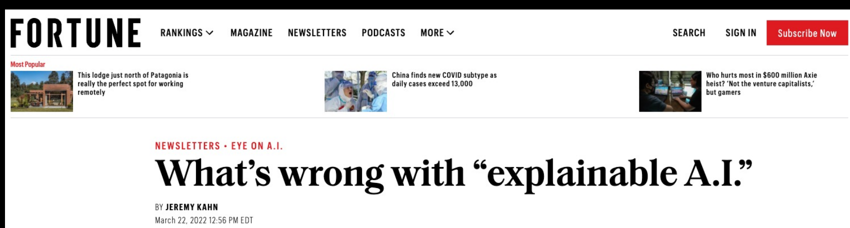
Why Trustworthy ML?

The New York Times

Account

The Secretive Company That Might End Privacy as We Know It

A little-known start-up helps law enforcement match photos of unknown people to their online images — and "might lead to a dystopian future or something," a backer says.



FORTUNE

RANKINGS MAGAZINE NEWSLETTERS PODCASTS MORE

SEARCH SIGN IN Subscribe Now

Most Popular

This lodge just north of Patagonia is really the perfect spot for working remotely

China finds new COVID subtype as daily cases exceed 13,000

Who hurts most in $600 million Axie heist? 'Not the venture capitalists,' but gamers

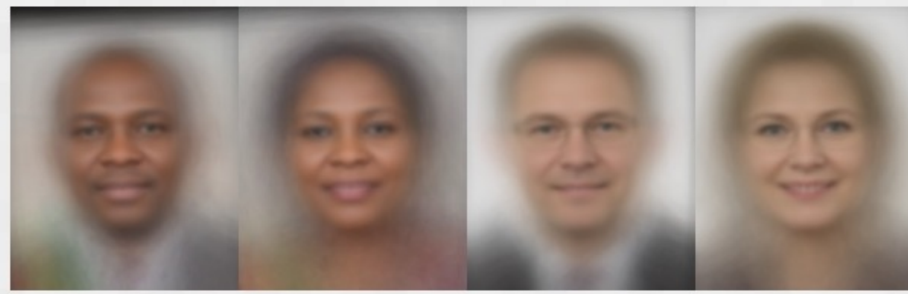
NEWSLETTERS • EYE ON A.I.

What's wrong with "explainable A.I."

BY JEREMY KAHN

March 22, 2022 12:56 PM EDT

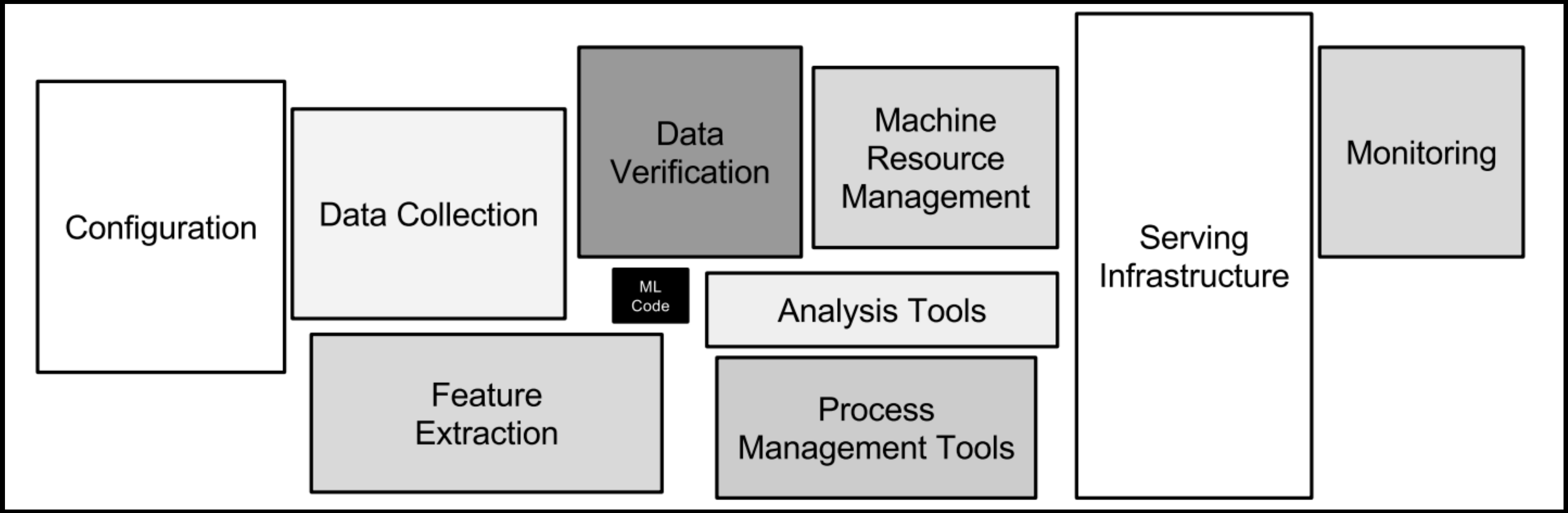
Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%



- Academic ML Research is "known", but industry is "unknown"
  - High-stake decision making: who should get bail? hired? a loan?
  - How is data collected?
  - How are ML systems evaluated?

[Noble 2018; Broussard 2018; Benjamin 2019; Gebru 2020; Benjamin 2020; Lakkaraju et al. 2020; Varshney 2022]

Let's talk about the "system": lessons from MLOps



[Sculley et al. 2015]

	ML Research	ML Production
Objective	Model performance	Different stakeholders == different objectives
Computational priority	Fast training; high throughput	Fast inference; low latency
Data	Static	Constant shifting
Fairness	Good to have (sadly)	Important
Interpretability	Good to have	Important

[Huyen 2022]

Agenda

1. Trustworthy ML
   a. Why do we care?

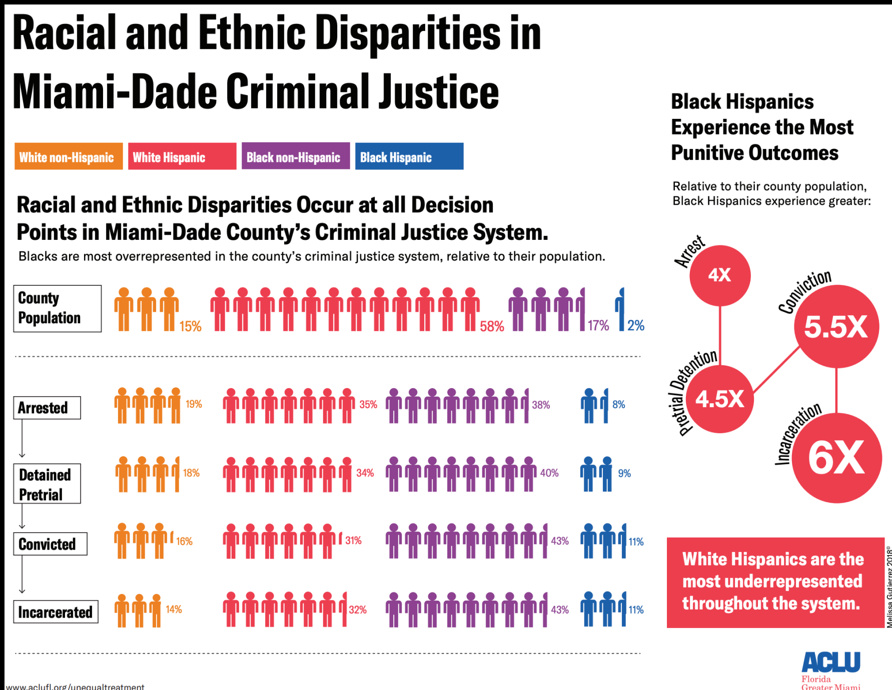
2. Problem: unequal racial treatment
   a. Social justice issues
   b. Vision and ML-based FPT issues

3. Equitable DL methodology
   a. Data and tackling biases
   b. Phase 1: multidimensionality of race
   c. Phase 2: “self-auditing” evaluation

4. Discussion & Recommendations

UNIVERSITY
OF MIAMI

Problem: Unequal Racial Treatment (Social Justice)



[Petersen et al. 2018]

Race-Ethnic Subgroup	U.S. General	MDC General	MDC Defendants
Black Hispanic	0.4%	1.9%	9.18%
White Hispanic	8.7%	58.4%	39.70%
Black non–Hispanic	12.2%	17.1%	37.96%
White non–Hispanic	63.7%	15.4%	13.14%
Total	**100.0%**	**100.0%**	**99.98%** *

* Other racial–ethnic groups represented a very small (0.02%) proportion and were removed from the dataset.

[Dass et al. 2020]

- Large-scale racial disparities in the U.S. criminal justice system [Ulmer 2012; Baumer 2013]
- True scope of systemic racial disparities masked due to missing race information [Fox and Swatt 2009; Grosso et al. 2014]

Related Social Justice Problems

- If CJ datasets contain race data, current methods to fill missing ethnicity labels:
  - Relying on text-based approach via the U.S. Hispanics Surnames List [Word and Perkins 1996; Wei et al. 2006; Word et al. 2008; Elliott et al. 2009; King and Johnson 2016]
  - Subjective human raters' assessments via visual inspection [Blair et al. 2004; King and Johnson 2016; Petersen 2017]

- How does race/ethnicity and facial-characteristics matter in criminal justice?
  - Features such as Afrocentric features, skin tone, etc.
  - Outcomes such as arrest, pre-trial, sentencing, incarceration

Agenda

1. Trustworthy ML
   a. Why do we care?

2. Problem: unequal racial treatment
   a. Social justice issues
   b. Vision and ML-based FPT issues

3. Equitable DL methodology
   a. Data and tackling biases
   b. Phase 1: multidimensionality of race
   c. Phase 2: “self-auditing” evaluation

4. Discussion & Recommendations

UNIVERSITY OF MIAMI

Some terminology…

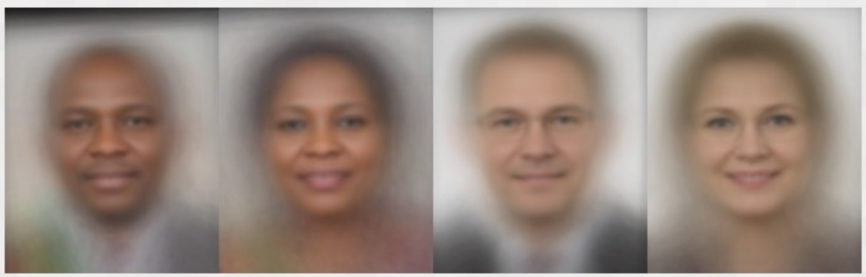
- Facial Processing Technology [Raji et al. 2020]:
  - Facial detection: localization and verification of a face
  - Facial segmentation: detection + alignment + cropping a face
  - Facial analysis: race, gender, age, facial landmarks, etc.
  - Face recognition: identification based on 1:1 and 1:N

In this dissertation project, race is studied as a facial analysis feature

- DL == Deep Learning
- DLM == Deep Learning Model

Problem: Unequal Racial Treatment (FPT)

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%



[Buolawamini and Gebru 2018]

Audits of commercial FPTs:

- Biased classifications against females and people of color [Buolawamini and Gebru 2018; Raji and Buolawamini 2019; Raji et al. 2020]

- Massive public and research community outcry have caused:
  - Bans and moratoria for the use of FPT across the world [Raji 2021]
  - Complete shutdown (IBM and Meta) and major overhaul (Microsoft and Amazon) of FPT-related projects [Smith 2018; Krishna 2020; Pesenti 2021]

Ongoing FPT debates

Cons

- Reinforces societal biases and worsen disparities in the CJ system
- Continue ignoring and recycling inherently flawed standard DL methods
- Trustworthiness and DL-based biases are considered “after-thought”

Pros

- Rich sociotechnical system – if responsibly developed can it address CJ disparities?
- Force ML community to rethink existing approaches and foster greater AI trust
- Shutting down its development **CANNOT** be the answer, many socially positive use cases:
  - Identify missing/trafficked children
  - Diagnosing hard to detect/rare diseases
  - Biometric security

Research Questions

- Identify and address different forms of harmful biases within an end-to-end DL classification pipeline
  - 4 types of biases: labeling bias; representation/data bias; algorithmic bias; evaluation bias [Suresh and Guttag 2021]
  - Distinct components: Data annotation and preprocessing; DLM training; DLM evaluation (inference and interpretation)

- Phase 1 – Multidimensionality of race
  - How is race considered in the vision literature?
  - Would a DLM's performance vary if the classification task changed from race to race/ethnicity prediction?

- Phase 2 – Create a rigorous evaluation strategy to assess:
  - DLM's inference performance per racial subgroup
  - Interpret DLM's performance: visualize what the DLM "sees"

Research Goals

- Collaborate with social science/CJ stakeholders throughout **entire process** of DLM development

- Create an equitable DL methodology for generating and interpreting racial categories using mugshots
  - **NOT** about a "typical" contribution to the literature
  - **Rethinking** existing standard approaches used in DL-based image classification based on "experimentation-based" approaches [Muthukumar et al. 2018; Balakrishnan et al. 2021]

- Provide empirical support and cautionary arguments for **the specific use** of the proposed DL methodology
  - Foster AI trustworthiness: rigorously assess an equitable FPT
  - Fill missing CJ race labels and uncover racial disparities at scale

Agenda

1. Trustworthy ML
   a. Why do we care?

2. Problem: unequal racial treatment
   a. Social justice issues
   b. Vision and ML-based FPT issues

3. Equitable DL methodology
   a. Data and tackling biases
   b. Phase 1: multidimensionality of race
   c. Phase 2: “self-auditing” evaluation

4. Discussion & Recommendations

UNIVERSITY
OF MIAMI

Data and Interdisciplinary Approaches (1/2)

- Analyzed a novel dataset of 195K MDC arrestees' mugshots (2010-2015)

- UM Sociology Student Raters Survey 14K stratified samples (29-labels) including:
  - Two Race (Black and White)
  - Four Race-Ethnicity (Black Hispanic, White Hispanic, Black Non-Hispanic, White Non-Hispanic)

- Tackle labeling bias:
  - Single-rater "court" labels
  - Consensus-rating "student" labels

Race-Ethnic Subgroup	U.S. General	MDC General	MDC Arrestees
Black Hispanic	0.4%	1.9%	9.18%
White Hispanic	8.7%	58.4%	39.70%
Black non–Hispanic	12.2%	17.1%	37.96%
White non–Hispanic	63.7%	15.4%	13.14%
Total	**100.0%**	**100.0%**	**99.98%***

* Other racial–ethnic groups represented a very small (0.02%) proportion and were removed from the dataset.

[Dass et al 2020]

Data and Interdisciplinary Methods (2/2)

Tackle data/representation bias:

- Sample size: Balanced vs. Imbalanced
- Face preprocessing: Original vs. OpenFace
- Additional API augmentations
- Randomized sampling + seed

(a) Raw Black Mugshot (b) Raw White Mugshot

(c) OpenFace Black Mugshot (d) OpenFace White Mugshot

[Dass et al. 2020]

Tackle algorithmic bias:

- 7 deep CNN architectures
  - Baseline: AlexNet and VGGs
  - Contemporary: (SE-)ResNe(X)ts
- ImageNet pretraining
- One-cycle and differential learning rates

Agenda

1. Trustworthy ML
    a. Why do we care?

2. Problem: unequal racial treatment
    a. Social justice issues
    b. Vision and ML-based FPT issues

3. Equitable DL methodology
    a. Data and tackling biases
    b. Phase 1: multidimensionality of race
    c. Phase 2: "self-auditing" evaluation

4. Discussion & Recommendations

Multidimensionality of Race

- Lack of research concerning Hispanic face classification within the Computer Vision, Sociolegal and Criminology communities

- In the CV literature, person's "race" is seen to belong to *one* of several categories White, Black, Hispanic, South Asian…

- From Critical Race Theory, "race" **SHOULD NOT** be considered as singular but a "*multidimensional*" construct, i.e. Black Hispanic or White non-Hispanic, etc. [Hanna et al. 2019]

Phase 1: Black and White Classification Results

Model	Raw Images		OpenFace	
	Courts	Students	Courts	Students
ResNet–50	92.00%	93.50%	93.73%	91.72%
AlexNet	92.00%	92.75%	92.73%	89.72%
Inception–v4	94.25%	92.00%	93.98%	88.22%
SE–ResNet–50	93.75%	93.50%	93.98%	91.47%
SE–ResNext–50_32x4d	93.75%	89.25%	94.23%	89.72%
VGG–16_bn	94.00%	92.25%	92.23%	**93.98%**
VGG–19_bn	94.25%	92.50%	**94.48%**	91.47%

(a) Balanced classification: 1,000 samples per race subgroup.

Model	Raw Images	OpenFace
	Courts	Courts
ResNet–50	97.20%	**97.21%**
AlexNet	97.17%	96.84%
Inception–v4	97.26%	96.79%
SE–ResNet–50	97.37%	97.18%
SE–ResNext–50_32x4d	97.52%	97.12%
VGG–16_bn	97.45%	97.13%
VGG–19_bn	97.50%	97.08%

(b) Imbalanced classification: full Miami–Dade County arrestee population.

[Dass et al. 2020]

- After 28-experiemnts, both sets of DLMs achieved greatest accuracies of 94.48% (courts) and 91.47% (students) after OpenFace Preprocessing
- No singular model architecture performed best under all experimental settings => validates experimentation-based approach!
- Imbalanced vs. balanced highest overall accuracies: ResNet50 (courts, OpenFace) gain of only 2.73% compared to VGG19 (courts, OpenFace) despite using approx. 100-times more data!

Phase 1: Four Race/Ethnicity Classification Results

Model	Raw Images		OpenFace	
	Courts	Students	Courts	Students
ResNet–50	56.20%	73.30%	55.31%	70.71%
AlexNet	58.75%	75.87%	60.95%	73.46%
Inception–v4	59.00%	71.25%	51.43%	67.83%
SE–ResNet–50	61.12%	76.25%	**61.32%**	**74.84%**
SE–ResNext–50_32x4d	61.25%	79.12%	48.31%	70.46%
VGG–16_bn	60.50%	76.37%	58.19%	74.09%
VGG–19_bn	63.87%	77.12%	59.57%	74.09%

(a) Four race–ethnicity classification: balanced (1,000) samples per race subgroup.

Model	Raw Images	OpenFace
	Courts	Courts
ResNet–50	80.60%	80.93%
AlexNet	79.09%	79.93%
Inception–v4	80.79%	80.18%
SE–ResNet–50	80.61%	**81.05%**
SE–ResNext–50_32x4d	80.40%	80.77%
VGG–16_bn	80.26%	77.92%
VGG–19_bn	80.43%	79.77%

(b) Four race–ethnicity classification: imbalanced full arrestee population.

[Dass et al. 2020]

- Student DLMs outperformed court DLMs by 12.51% to 22.15%
- Average balanced court, OpenFace DLMs - 56.44% – not helpful!
- SE-ResNet50 singularly outperformed for OpenFace data

Phase 1: Four Race/Ethnicity Classification Results

Model	Raw Images		OpenFace	
	Courts	Students	Courts	Students
SE–ResNet–50	61.12%	76.25%	**61.32%**	**74.84%**

(a) Four race–ethnicity classification: balanced (1,000) samples per race subgroup.

Model	Raw Images	OpenFace
	Courts	Courts
SE–ResNet–50	80.61%	**81.05%**

(b) Four race–ethnicity classification: imbalanced full arrestee population.

[Dass et al. 2020]

Limitations and future improvements:

- “highest” performance: Imbalanced (81.05%) vs. Balanced (61.32%):
  - Improved by 19.74%
  - But due to 50-times more data
  - Suspicious as WH and BnH represent 75% of data

Next steps:

- Go beyond reporting “population” test accuracies
- Look into DLM performance for individual racial subgroups
- DLMs are complex! “Post-hoc” interpretable methods?

Agenda

1. Trustworthy ML
   a. Why do we care?
   b. Multidisciplinary perspectives

2. Problem: unequal racial treatment
   a. Social justice issues
   b. Vision and ML-based FPT issues

3. Equitable DL methodology
   a. Data and tackling biases
   b. Phase 1: multidimensionality of race
   c. Phase 2: "self-auditing" evaluation

4. Discussion & Recommendations

UNIVERSITY OF MIAMI

Phase 2: Interdisciplinary Results (1/8)

- Generated approx. 194K mugshots
- High degree of correspondence with generated court labels, r = 0.8143
- Suggests a viable method for generating missing race-ethnicity labels in court databases
- Expand to investigate disparities in criminal justice

```
$ cat mtCt2s_vgg19.csv | head -10
Jail_ID,Black_Prob,White_Prob,Prediction
110068060,0.99995,5e-05,Black
120000006,0.24984,0.75016,White
120000034,0.99997,3e-05,Black
120000104,0.9838,0.0162,Black
120000109,0.00037,0.99963,White
120000164,0.99998,2e-05,Black
120000171,0.99997,3e-05,Black
120000182,0.99999,1e-05,Black
120000195,1.0,0.0,Black
```

```
rdass@sickles ~/Research/ULINK/2020_Image_Vision_Computing/cat4_full_dataset
$ nvidia-smi
Mon Jun  1 00:01:40 2020
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 384.81                 Driver Version: 384.81                    |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  Tesla P100-PCIE...  Off  | 00000000:04:00.0 Off |                    0 |
| N/A   44C    P0    43W / 250W |  14074MiB / 16276MiB |     12%      Default |
+-------------------------------+----------------------+----------------------+
|   1  Tesla P100-PCIE...  Off  | 00000000:82:00.0 Off |                    0 |
| N/A   40C    P0    37W / 250W |  14145MiB / 16276MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                       GPU Memory |
|  GPU       PID   Type   Process name                             Usage      |
|=============================================================================|
|    0     23987      C   ...rdass/Research/tensorflowEnv/bin/python   723MiB |
|    0     23995      C   ...rdass/Research/tensorflowEnv/bin/python  1957MiB |
|    0     26953      C   ...rdass/Research/tensorflowEnv/bin/python   723MiB |
|    0     26973      C   ...rdass/Research/tensorflowEnv/bin/python  1947MiB |
|    0     27275      C   ...rdass/Research/tensorflowEnv/bin/python   723MiB |
|    0     27315      C   ...rdass/Research/tensorflowEnv/bin/python   723MiB |
|    0     31480      C   ...rdass/Research/tensorflowEnv/bin/python   723MiB |
|    0     31799      C   python                                      1279MiB |
|    0     36250      C   ...rdass/Research/tensorflowEnv/bin/python  1945MiB |
|    0     36519      C   ...rdass/Research/tensorflowEnv/bin/python  3299MiB |
|    1     23987      C   ...rdass/Research/tensorflowEnv/bin/python  2145MiB |
|    1     26953      C   ...rdass/Research/tensorflowEnv/bin/python  2161MiB |
|    1     27275      C   ...rdass/Research/tensorflowEnv/bin/python  3259MiB |
|    1     27315      C   ...rdass/Research/tensorflowEnv/bin/python  3299MiB |
|    1     31480      C   ...rdass/Research/tensorflowEnv/bin/python  3259MiB |
+-----------------------------------------------------------------------------+
```

Phase 2: Extension of Phase 1 Methods

- Contemporary Vision architectures (ImageNet benchmarks)
  - DenseNet121
  - Unable to load (RL-based) NAS and PNAS architectures on Sickles!

- Extended face preprocessing:
  - Original vs. OpenFace vs. MTCNN

- Proposing "self-auditing" strategy for disaggregated evaluation

Phase 2: Results (2/8) – Extent of Face Preprocessing

Original [varying resolution]	Original resized [299 x 299]	OpenFace [299 x 299]	MTCNN [299 x 299]

[N = 195,174]

[N = 194,957]
-217

[N = 195,162]
-12

[Dass et al. 2022]

Phase 2: Binary Classification Validation Results (3/8)

Model	original		OpenFace		MTCNN	
	Court	Students	Court	Students	Court	Students
AlexNet	92.50%	**94.00%**	97.25%	92.25%	92.75%	93.00%
DenseNet161	**97.00%**	93.00%	97.00%	92.25%	96.50%	94.00%
InceptionV4	93.25%	90.25%	90.75%	90.50%	90.75%	91.75%
ResNet50	96.50%	92.50%	97.00%	91.50%	95.25%	93.25%
SE-ResNet50	95.00%	92.75%	**97.75%**	**92.50%**	96.75%	93.25%
SE-ResNeXt50	96.75%	91.75%	97.75%	90.00%	96.75%	94.25%
VGG19	96.00%	89.75%	97.00%	92.25%	**96.75%**	**94.75%**

[Dass et al. 2022]

- Court-labeled, OpenFace-preprocessed SE-ResNet50 model (97.75%)
  - Optimal model experimental combination
- Experimentation-based approach to tackle “No Free Lunch Theorem”
  - AlexNet (94%) highest accuracy for student-labeled, original data
  - Cannot assume that “best ImageNet architecture” would be optimal for our task

Fairness metric: Disaggregated Evaluation (1/2)

- Rather than assessing overall DLM's performance as a population (Phase 1) – i.e., Black *and* White [Mitchel et al. 2019]
- Tackle evaluation bias: segregate test datasets and report its performance on individual subgroups – i.e., Black *or* White
- Based on six stratified datasets, segregating them based on race => 12 test datasets
  - Each with unique test augmentation parameters
  - Each with unique test sample size
- For DLM assessment: keep labeling source constant in terms of training and testing data
  - Court-trained DLM will be only tested on court-annotated data

Fairness metric: Disaggregated Evaluation (2/2)

Test dataset	Test dataset size	Test augmentation parameters		
		Ground-truth	Face preprocessing	Target Racial category
1	5,931	court	original	Black
2	6,244	court	original	White
3	5,924	court	OpenFace	Black
4	6,242	court	OpenFace	White
5	5,931	court	MTCNN	Black
6	6,244	court	MTCNN	White
7	6,198	student	original	Black
8	5,818	student	original	White
9	6,190	student	OpenFace	Black
10	5,817	student	OpenFace	White
11	6,198	student	MTCNN	Black
12	5,818	student	MTCNN	White

[Dass et al. 2022]

Phase 2: "Self-auditing" method

252 model inference interpretability scenarios
*= **42** DLMs and **12** "experimental parameters"*
(unseen mugshots from same dataset)

- Inference: predict binary (Black vs. White) racial categories
  - Ground-truth source: Courts vs. Students
  - Extent of face preprocessing: Original vs. OpenFace vs. MTCNN
  - Racial category: Black vs. White

- Interpretability "post-hoc" method: visualize DLM top-layer
  - Saliency maps: Grad-CAM and guided backpropagation
  - Greatest model confidence: correctly (best) and incorrectly (worst) mugshots (DLM blind spots)

Phase 2: Results (4/8) - Self-auditing Court DLMs

5 **Highest** accuracies for unseen Black and White mugshots

Training data	Training architecture	Test dataset	Test race label	Test accuracy
OpenFace	DenseNet161	1	Black	99.92%
MTCNN	SE-ResNet50	1	Black	99.65%
OpenFace	SE-ResNet50	1	Black	99.49%
MTCNN	InceptionV4	1	Black	99.41%
OpenFace	ResNet50	1	Black	99.09%

Training data	Training architecture	Test dataset	Test race label	Test accuracy
original	SE-ResNeXt50	4	White	99.60%
MTCNN	VGG19	4	White	99.52%
MTCNN	DenseNet161	4	White	98.77%
OpenFace	InceptionV4	2	White	98.51%
original	AlexNet	4	White	98.46%

5 **Lowest** accuracies for unseen Black and White mugshots

Training data	Training architecture	Test dataset	Test race label	Test accuracy
OpenFace	InceptionV4	1	Black	2.56%
original	SE-ResNeXt50	3	Black	58.90%
original	AlexNet	3	Black	67.23%
MTCNN	VGG19	3	Black	71.29%
original	ResNet50	3	Black	72.57%

Training data	Training architecture	Test dataset	Test race label	Test accuracy
MTCNN	InceptionV4	2	White	1.28%
OpenFace	DenseNet161	2	White	6.13%
OpenFace	ResNet50	2	White	21.68%
OpenFace	SE-ResNet50	2	White	22.90%
MTCNN	AlexNet	2	White	24.60%

[Dass et al. 2022]

Phase 2: Results (5/8) - Self-auditing Student DLMs

5 **Highest** accuracies for unseen Black and White mugshots

Training data	Training architecture	Test dataset	Test race label	Test accuracy
MTCNN	InceptionV4	7	Black	99.58%
MTCNN	DenseNet161	7	Black	99.40%
MTCNN	SE-ResNet50	7	Black	99.14%
MTCNN	VGG19	7	Black	98.24%
OpenFace	DenseNet161	9	Black	97.71%

Training data	Training Architecture	Test dataset	Test race label	Test accuracy
original	AlexNet	10	White	98.26%
original	ResNet50	10	White	97.85%
MTCNN	ResNet50	10	White	96.72%
MTCNN	SE-ResNeXt50	12	White	96.55%
MTCNN	AlexNet	10	White	95.53%

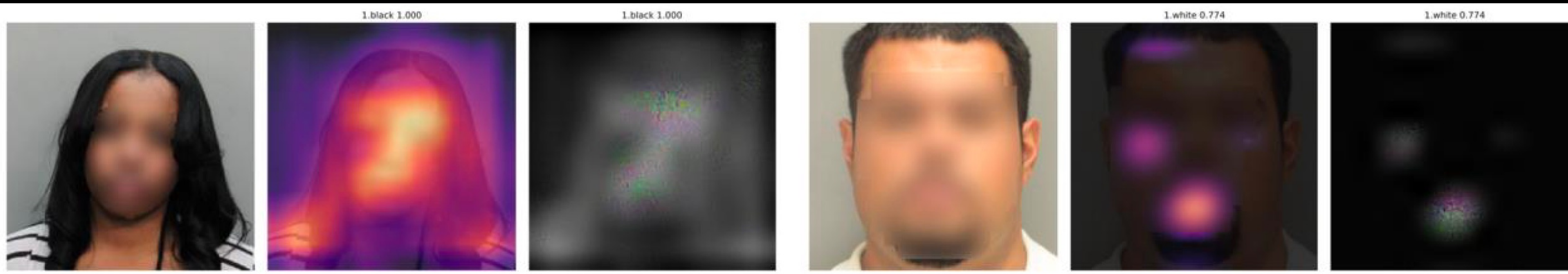
5 **Lowest** accuracies for unseen Black and White mugshots

Training data	Training architecture	Test dataset	Test race label	Test accuracy
OpenFace	VGG19	7	Black	52.40%
OpenFace	SE-ResNet50	7	Black	61.67%
original	AlexNet	9	Black	71.83%
original	ResNet50	9	Black	77.37%
OpenFace	AlexNet	7	Black	82.22%

Training data	Training architecture	Test dataset	Test race label	Test accuracy
OpenFace	InceptionV4	8	White	5.38%
MTCNN	DenseNet161	8	White	23.22%
MTCNN	InceptionV4	8	White	28.24%
MTCNN	AlexNet	8	White	37.71%
MTCNN	ResNet50	8	White	49.72%

[Dass et al. 2022]

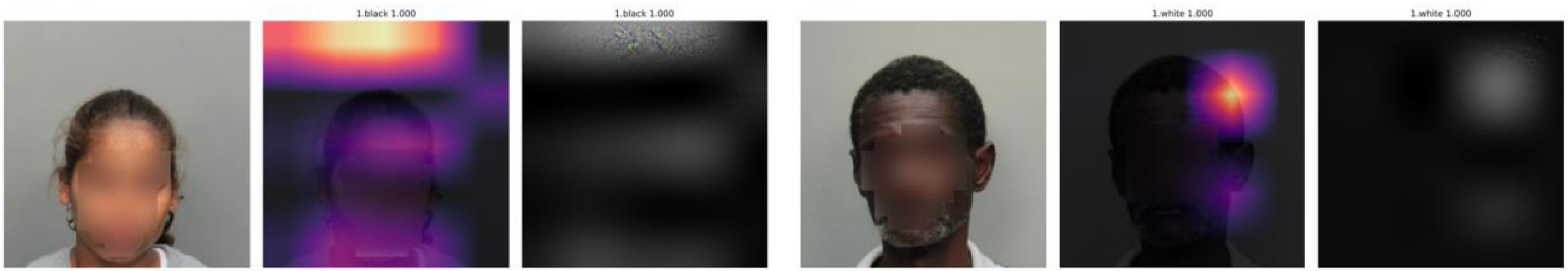
Comparing Court DLM "post-hoc" Results (6/8)



(a) "Best" Black mugshot by OpenFace preprocessed court trained DenseNet161 model.

(b) "Worst" Black mugshot by OpenFace preprocessed court trained DenseNet161 model.

Model with the highest test accuracy (99.92%) for Black mugshots



(e) "Best" Black mugshot by OpenFace preprocessed court trained InceptionV4 model.

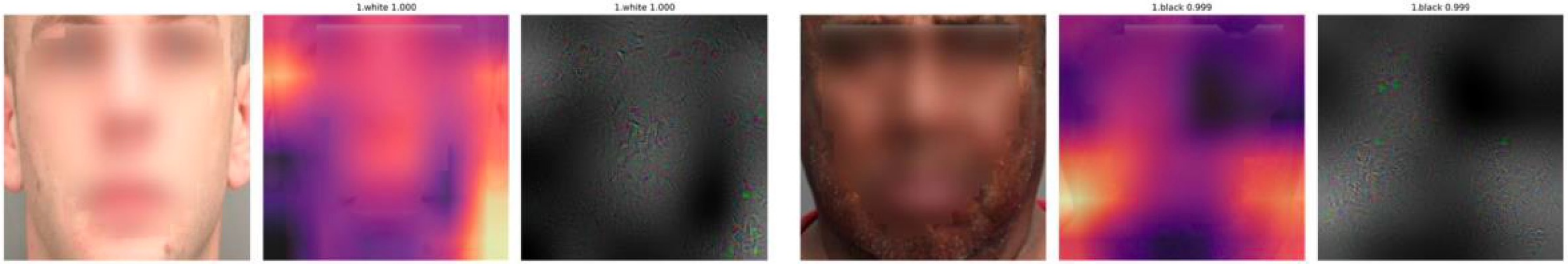
(f) "Worst" Black mugshot by OpenFace preprocessed court trained InceptionV4 model.

Model with the lowest test accuracy (2.56%) for Black mugshots

Training data	Training architecture	Model validation accuracy	Test dataset	Test race label	Model testing accuracy	"Best" mugshot confidence	"Worst" mugshot confidence
OpenFace	DenseNet161	97.00%	1	**Black**	**99.92%**	100%	77.4%
OpenFace	InceptionV4	90.75%	1	**Black**	**2.56%**	100%	100%

[Dass et al. 2022]

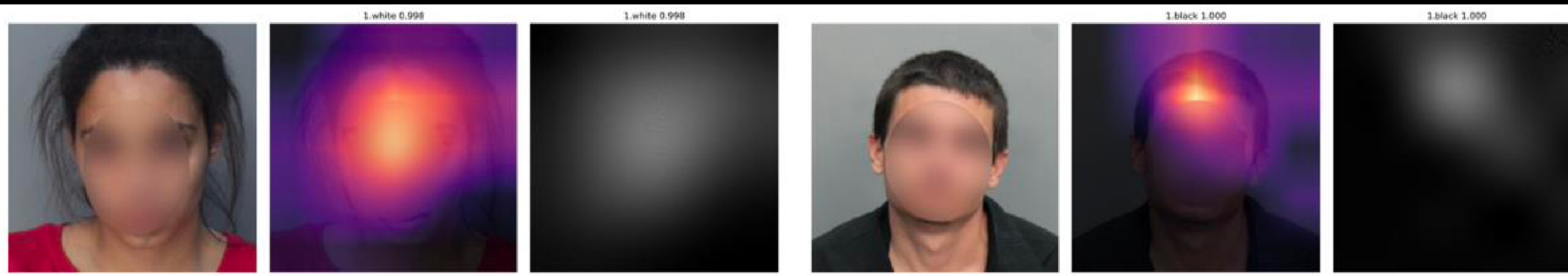
Comparing Student DLM “post-hoc” Results (5/5)



(a) “Best” White mugshot by original resized student trained AlexNet model.

(b) “Worst” White mugshot by original resized student trained AlexNet model.

Model with the **highest** test accuracy (98.26%) for White mugshots



(e) “Best” White mugshot by MTCNN preprocessed student trained InceptionV4 model.

(f) “Worst” White mugshot by MTCNN preprocessed student trained InceptionV4 model.

Model with the **lowest** test accuracy (5.38%) for White mugshots

Training data	Training architecture	Model validation accuracy	Test dataset	Test race label	Model testing accuracy	“Best” mugshot confidence	“Worst” mugshot confidence
original	AlexNet	94.00%	10	**White**	**98.26%**	100%	99.9%
OpenFace	InceptionV4	90.50%	8	**White**	**5.38%**	99.8%	100%

[Dass et al. 2022]

Agenda

1. Trustworthy ML
    a. Why do we care?
    b. Multidisciplinary perspectives

2. Problem: unequal racial treatment
    a. Social justice issues
    b. Vision and ML-based FPT issues

3. Equitable DL methodology
    a. Data and tackling biases
    b. Phase 1: multidimensionality of race
    c. Phase 2: "self-auditing" evaluation

4. Discussion & Recommendations

UNIVERSITY OF MIAMI

Discussion and Recommendations (1/2)

Test racial category	Court		Student	
	Ten most accurate (average test accuracy)	**Ten least accurate** (average test accuracy)	**Ten most accurate** (average test accuracy)	**Ten least accurate** (average test accuracy)
Black	98.55%	66.44%	98.17%	76.79%
White	98.33%	32.17%	96.50%	45.39%
Difference (gain for Black race)	**0.22%**	**34.27%**	**1.67%**	**31.40%**

[Dass et al. 2022]

- Across 80 "most impactful" inference disaggregated cases:
  - 40 highest + 40 lowest test accuracies for Black and White mugshots
- On average, test accuracies for Black mugshots consistently outperformed White mugshots by 0.22% to 34.27%
- Surprisingly contradicts "Gender Shades" findings + **DOES NOT** perpetuate current notions of "embedded" bias

Discussion and Recommendations (2/2)

- Strong evidence for model robustness:
  - Overwhelming majority of 80 scenarios, face preprocessing method applied during training is *different* from the inference dataset

- Self-Auditing interpretation results:
  - 32 “best” and “worst” mugshots – both saliency maps largely focus on the face
    - Black mugshots: lower nasal and mouth
    - White mugshots: upper cheekbone, mid-nasal and forehead
  - Overall, they are inconsistent, however this is good thing:
    - Opposes notions that DLMs are biased w.r.t race
    - Accuracies alone DO NOT reveal the whole picture
    - Valuable insights to better understand DLM generalizability

Ethics and Project limitations

Ethical considerations

- Define project scope: fill missing race CJ data to help uncover CJ racial disparities
- Protect individuals' privacy: blur mugshots for research/public dissemination
- Provide full transparency regarding end-to-end DLM pipeline:
  - Data collection and annotation; DLM training; DLM evaluation and interpretation
  - Open to providing trained DLM weights but will not share raw mugshot data

Limitations

- Easy classification task ➔ High model accuracies?
- New biases: racial categories (sampling bias); MDC raters (labeling bias)
- Skin tone or facial features proxy/correlated to race?
- Experimentation-based methods results ➔ scalability issue
  - 1,000+ cases to analyze just for binary race!
- Trying to make "fairer" ML systems ➔ might be illegal
  - Explicitly considering race as a classification task is illegal under Equal Credit Opportunity Act
  - Bu, if we ignore race ➔ society and ML-based systems are increasingly "color-blind" [Bonilla-Silva 2006]

Future Work

- Investigate intersectional disaggregated evaluation via self-auditing, i.e., race-ethnicity [Mitchel et al. 2019]

- Extend methodology to other CJ databases or other face benchmark datasets (FairFace; VMER; etc.)

- Modify the DLM classification task to other attributes such as gender or skin tone to understand other systemic disparities in CJ

- Investigate effects of inverted faces (Thatcher effect) and other model initializations paradigms (ImageNet vs. Random vs. face pretrained)

- Disaggregated evaluation may not account for randomness within inference datasets, consider other metrics such as confidence intervals and p-values that considers uncertainties [Barocas et al. 2021]

Final Thoughts and Conclusion

- To foster greater AI trustworthiness:
  - Bring (domain specific) “trustworthy” elements to the forefront of product design, development and evaluation
  - Include target domain experts and their insights throughout the entire process

- Using experimentation-based approaches, developed an equitable DL methodology within an FPT sociotechnical (Vision-based) system for generating and interpreting racial categories using mugshots
  - Mitigated 4 types of biases within separate components in a DL pipeline
  - Considering race as multidimensional is difficult even for DLMs
  - Proposed a “self-auditing” strategy for disaggregated evaluation
    - Critical finding: DLMs predicted Black mugshots with higher accuracies than White counter parts by 0.22% to 34.27%
    - Human in the loop + “Post-hoc” methods is essential, if DL systems deployed in high-stakes decision making domains

References

Noble, S. U. (2018). "Algorithms of oppression." New York University Press

Broussard, M. (2018). "Artificial unintelligence: How computers misunderstand the world." MIT Press

Benjamin, R. (2019). "Race after technology: Abolitionist tools for the new Jim code." *Social forces*.

Gebru, T. (2020). Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 3609-3609).

Benjamin, R. (2020). "The New Jim Code (Chapter)." *Which side of History? How Technology is Reshaping Democracy and Our Lives. Common Sense media.*

Lakkaraju, H. et al. (2020). "Explaining Machine Learning Predictions: State-of-the-art, Challenges, and Opportunities" NeurIPS Tutorial

Varshney, K. R. (2022). "Trustworthy Machine Learning." *Chappaqua, NY, USA: Independently Published*.

Sculley, D. et al. (2015). "Hidden technical debt in machine learning systems." *Advances in neural information processing systems*, *28*.

Huyen, C. (2022). "Designing Machine Learning Systems." *O'Reilly Media, Inc.*

Petersen, N. et al. (2018). "Unequal treatment: Racial and ethnic disparities in Miami-Dade criminal justice." *ACLU of Florida.*

Dass, R. K. et al. (2020). "It's not just black and white: classifying defendant mugshots based on the multidimensionality of race and ethnicity." In *2020 17th Conference on Computer and Robot Vision (CRV)* (pp. 238-245). IEEE.

Ulmer, J. T. (2012). "Recent developments and new directions in sentencing research." *Justice Quarterly*, *29*(1), 1-40.

Baumer, E. P. (2013). "Reassessing and redirecting research on race and sentencing." *Justice Quarterly*, *30*(2), 231-261.

Fox, J. A. and Swatt, M. L. (2009). "Multiple imputation of the supplementary homicide reports, 1976–2005." *Journal of Quantitative Criminology*, *25*(1), 51-77.

Grosso, C. M. et al. (2014). "Race discrimination and the death penalty: An empirical and legal overview." *America's experiment with capital punishment: Reflections on the past, present, and future of the ultimate penal sanction*, 525-576.

Word, D. L. and Perkins, R. C. (1996). "Building a Spanish Surname List for the 1990's--: A New Approach to an Old Problem." Washington, DC: Population Division, US Bureau of the Census.

Wei, I. I. el al. (2006). "Using a Spanish surname match to improve identification of Hispanic women in Medicare administrative data." *Health Services Research*, *41*(4p1), 1469-1481.

Word, D. L. et al. (2008). "Demographic aspects of surnames from census 2000." *Unpublished manuscript*.

Elliott, M. N. et al. (2009). "Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities." *Health Services and Outcomes Research Methodology*, *9*(2), 69-83.

King, R. D. and Johnson, B. D. (2016). "A punishing look: Skin tone and Afrocentric features in the halls of justice." *American Journal of Sociology*, *122*(1), 90-124.

Blair, I. V. et al. (2004). "The influence of Afrocentric facial features in criminal sentencing." *Psychological science*, *15*(10), 674-679.

Petersen, A. M. (2017). "Complicating race: Afrocentric facial feature bias and prison sentencing in Oregon." *Race and Justice*, *7*(1), 59-86.

Raji, I. D. et al. (2020). "Saving face: Investigating the ethical concerns of facial recognition auditing." In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 145-151).

References

Buolamwini, J. and Gebru, T. (2018). “Gender shades: Intersectional accuracy disparities in commercial gender classification.” In *Conference on fairness, accountability and transparency* (pp. 77-91). PMLR.

Raji, I. D. and Buolamwini, J. (2019). “Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products.” In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 429-435).

Raji, I. D. and Fried, G. (2021). “About Face: A survey of facial recognition evaluation.” *arXiv preprint arXiv:2102.00813*.

Smith, B. (2018). “Facial recognition technology: The need for public regulation and corporate responsibility.” *Microsoft on the Issues*.

Krishna, A. (2020). “IBM CEO’s Letter to Congress on Racial Justice Reform.” *THINKPolicy Blog, June*, *8*.

Pesenti, J. (2021). “Facebook: An update on our use of face recognition” *Meta Newsroom.*

Suresh, H. and Guttag, J. (2021). “A framework for understanding sources of harm throughout the machine learning life cycle.” In *Equity and access in algorithms, mechanisms, and optimization* (pp. 1-9).

Muthukumar, V. et al. (2018). “Understanding unequal gender classification accuracy from face images.” *arXiv preprint arXiv:1812.00099*.

Balakrishnan, G. et al. (2021). “Towards causal benchmarking of bias in face analysis algorithms.” *In Deep Learning-Based Face Analytics* (pp. 327-359). Springer, Cham.

Hanna, A. et al. (2020). “Towards a critical race methodology in algorithmic fairness.” In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 501-512).

Dass, R. K. et al. (2022). “Detecting racial inequalities in criminal justice: towards an equitable deep learning approach for generating and interpreting racial categories using mugshots.” *AI & SOCIETY*, 1-22.

Mitchell, M. et al. (2019). “Model cards for model reporting.” In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 220-229).

Bonilla-Silva, E. (2006). “Racism without racists: Color-blind racism and the persistence of racial inequality in the United States.” *Rowman & Littlefield Publishers.*

Barocas, S. et al. (2021). “Designing disaggregated evaluations of ai systems: Choices, considerations, and tradeoffs.” In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 368-378).

My next chapter…

Applied Scientist Intern @ Amazon

Fairness and Responsible AI group

Why Trustworthy ML?

"Trustworthiness begins with people, not AI, and what do we want from people who are trustworthy?"

$$\text{Org. Management}\begin{cases}\text{Confidence}\\ \text{Reliability}\\ \text{Openness}\\ \text{Selflessness}\end{cases} \qquad \text{ML systems}\begin{cases}\text{Accuracy}\\ \text{Robustness}\\ \text{Communication}\\ \text{Empowerment}\end{cases}$$

"Move beyond local task-specific optimizations and think global scaling issues", and,
"Epistemic uncertainty: ML outcomes that have nothing to do with probabilities"
Kush Varshney (Distinguished Scientist, IBM Research)

[NIST 2021 – AI Risk Management Workshop]

Impact of FATE Research

Fairness, **A**ccountability, **T**ransparency and **E**thics

- 750% increase in accepted papers (2017-2020) [Qian et al. 2021]
- FAccT, AIES “exclusive” conferences
- Since 2012 – full paper and workshop tracks
  - Vision (ICCV, CVPR)
  - AI/ML conferences (NeurIPS, ICML, AAAI, ICLR)
  - Robotics, Medical, NLP etc.
- ML journals
  - SI “AI for People”, 2022 AI & Society [Dass et al. 2022]
  - SI “Safe and Fair ML” 2022 Machine Learning
  - SI “Bias and Fair ML” 2021 Data Mining and Knowledge Discovery
- Industry research groups – PAIR (Google); FATE (Microsoft); FAIR (Facebook); AI Fairness 360 (IBM); E&S (Deepmind)

My Interdisciplinary Dissertation in a Nutshell (1/2)

- Continued SOTA progress with ML systems, but increased distrust by various stakeholders (researchers, public, etc.)
- Long-standing but timely issue of unequal treatment based on race – society and technology (sociotechnical)
- Facial Processing Technology – a tool exacerbating racial inequalities in CJ or used to help ameliorate them?
- 3+ year journey collaborating with social sciences/CJ domain experts studying the Miami-Dade County CJ system

My Interdisciplinary Dissertation in a Nutshell (2/2)

- Re-think standard approaches in end-to-end supervised DL image classification

- Proposing experimentation-based methods:
  - Tackle fairness and bias issues across different DL components
  - Race as multidimensional construct
  - Rigorous “self-auditing” evaluation approach:
    - Model Inference
    - Model Interpretation

- Offer empirical support + cautionary recommendations to ML/CJ stakeholders via an equitable FPT methodology