Investigating social bias in automated face classification at its roots:
Human annotator biases are learned and reproduced by face classification models

Qianqian(Lee) Cui, Jeffrey J. Berg, and David M. Amodio
NYU Social Neuroscience Lab

Introduction

Background

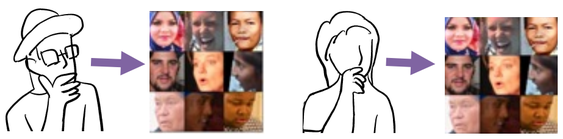
The potential for, and instances of, bias in artificial intelligence (AI) systems have been widely documented and discussed, with far-reaching societal implications. This is especially true regarding face classification: the ability for AI systems to predict the demographics of an individual from their face.

Nevertheless, the sources of bias in automated face classification systems remain poorly understood. When constructing face classification datasets, standard procedure involves recruiting a handful of human participants to manually label the demographic attributes of large numbers of face images.

To extent that the classifications rendered by the human annotators are biased, a classification model trained on their labels may learn and subsequently reproduce those biased judgments.

Research Question

Do AI-based classification models learn and reproduce the social biases of the human annotators who provide their training data?



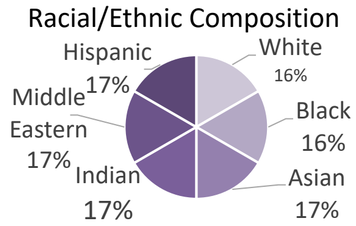
Hypotheses

- **H1**) Models trained on the (biased) labels provided by more prejudiced annotators will be less accurate in classifying non-White faces than models trained on the labels provided by less prejudiced annotators
- **H2**) Models trained on labels provided by more prejudiced annotators will show patterns of racial hypodescent when classifying racially ambiguous faces

Method

1. Collect large-scale face image dataset

- 15,000 faces balanced in terms of racial/ethnic composition



2. Human race/sex annotation

Participants:

600 Amazon Mechanical Turk workers from United States

Procedure:

1. Annotation task

Each participant will be randomly assigned to annotate (label) 200 faces on race/ethnicity and sex

2. Black/White – Good/Bad Implicit Association Task (IAT)

3. Prejudice-Related Self-Report Questionnaires

Feeling Thermometer	Racial Identity Strength
Modern Racism Scale	Perceived Racial Status Threat
Blatant Dehumanization	Racial Resentment
Social Dominance Scale	…

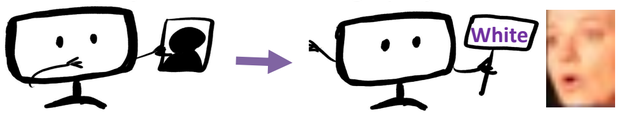
3. Training of face classification models

1. Divide annotations based on prejudice measure

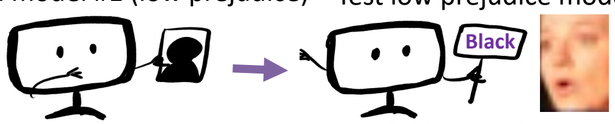
Low prejudice group: Modern Racism Score ≤ Median
High prejudice group: Modern Racism Score > Median

2. Train separate face classification models

Low prejudice model vs. high prejudice model



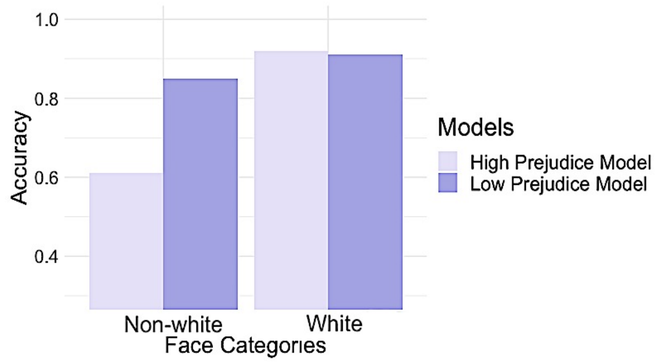
Train model #1 (low prejudice) Test low prejudice model



Train model #2 (high prejudice) Test high prejudice model

Predicted Classification Results

1. Accuracy on monoracial faces



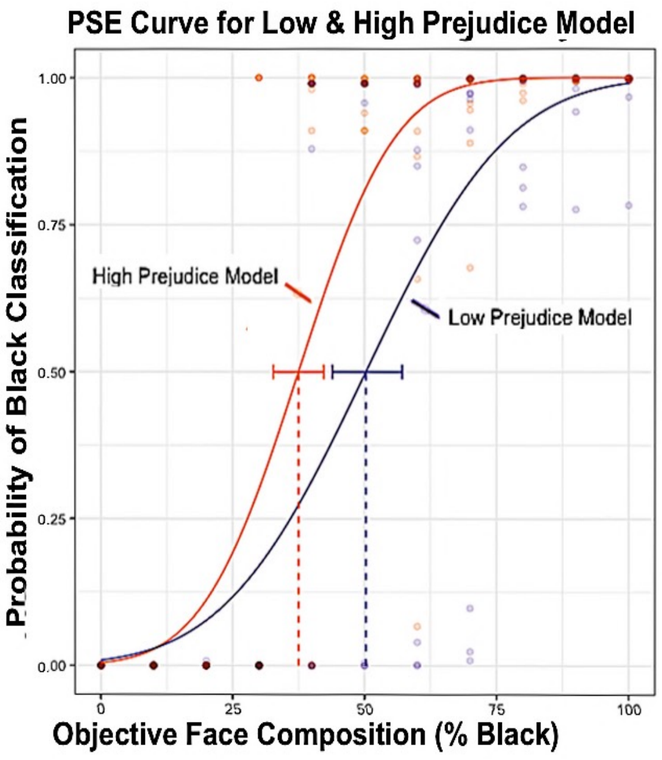
2. Classifications of ambiguous faces

Create Black-White face morphs

- 0% Black – 100% Black, 200 in total

Compute PSE for each model:

- *Point of subjective equality* (PSE): The point at which a face is equally likely to be categorized as Black or White (ranging from 0 to 1)



Discussion

Summary of Expectations

- Models trained on race labels provided by high-prejudice annotators will have lower (biased) accuracy rates for non-White faces, relative to models trained on labels provided by low-prejudice annotators
- Models trained on race labels provided by high-prejudice annotators will show a hypodescent-like pattern (PSE < .5) when classifying racially ambiguous faces
- Models trained on race labels provided by low-prejudice annotators will produce PSEs near .5

Additional Research Plans

- Collect and test both low-prejudice and high-prejudice models on dataset of real mixed-race individuals
- Explore bias in sex classifications, as well as race x sex interactions
- Explore relationships between measures of prejudice
- Use dimension reduction techniques to derive latent "prejudice" factor