Bias in Facial Recognition

George Wei, Bradon Thymes, Gudrun Thorkelsdottir, Tiye Kurtz-Miott, Ryan Downing, Femi Obiwumi, JJ Shankar, Rachel Mattson

Mentors: Prof. Tom Goldstein, Prof. John Dickerson, Micah Goldblum, Arpit Bansal, Samuel Dooley

COMPUTER SCIENCE
UNIVERSITY OF MARYLAND

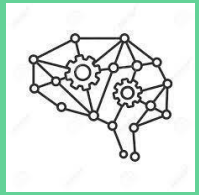
REU-CAAR 2021

Research Experiences For Undergraduates

Facial Recognition Systems

- There are two very effective types of facial recognition systems
  - Deep learning models
  - You!
- An example facial recognition task

Input



FR system subjects image to learned ideas about faces

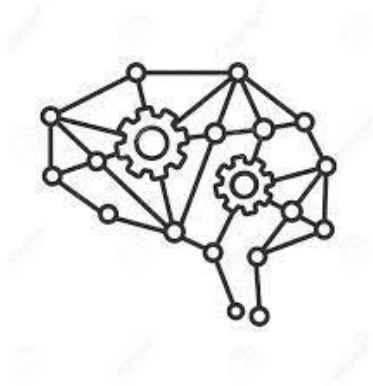
“Dr. Tom Goldstein”

Output

Project goals

1. Compare areas of weaknesses/bias between humans and A.I.
   a. Design an empirical study

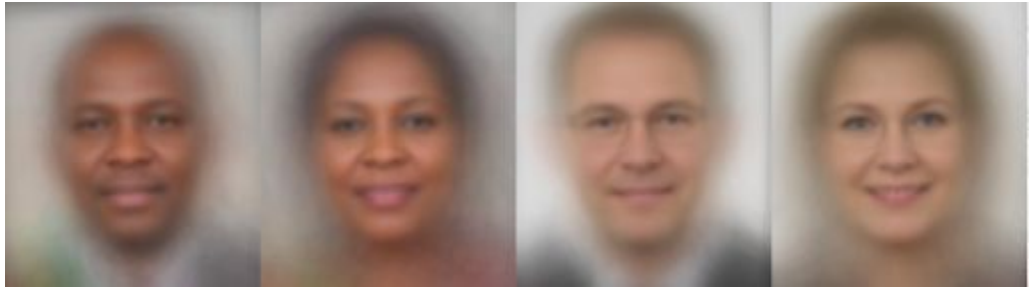
vs.



2. Along the way, identify issues with datasets that might lead to bias in A.I. systems, and create a clean dataset

Bias in Machine Learning Systems

- Gender Shades
- Discrimination in Online Ad Delivery
- Robustness Disparities in Commercial Face Detection



Bias in Human Recognition

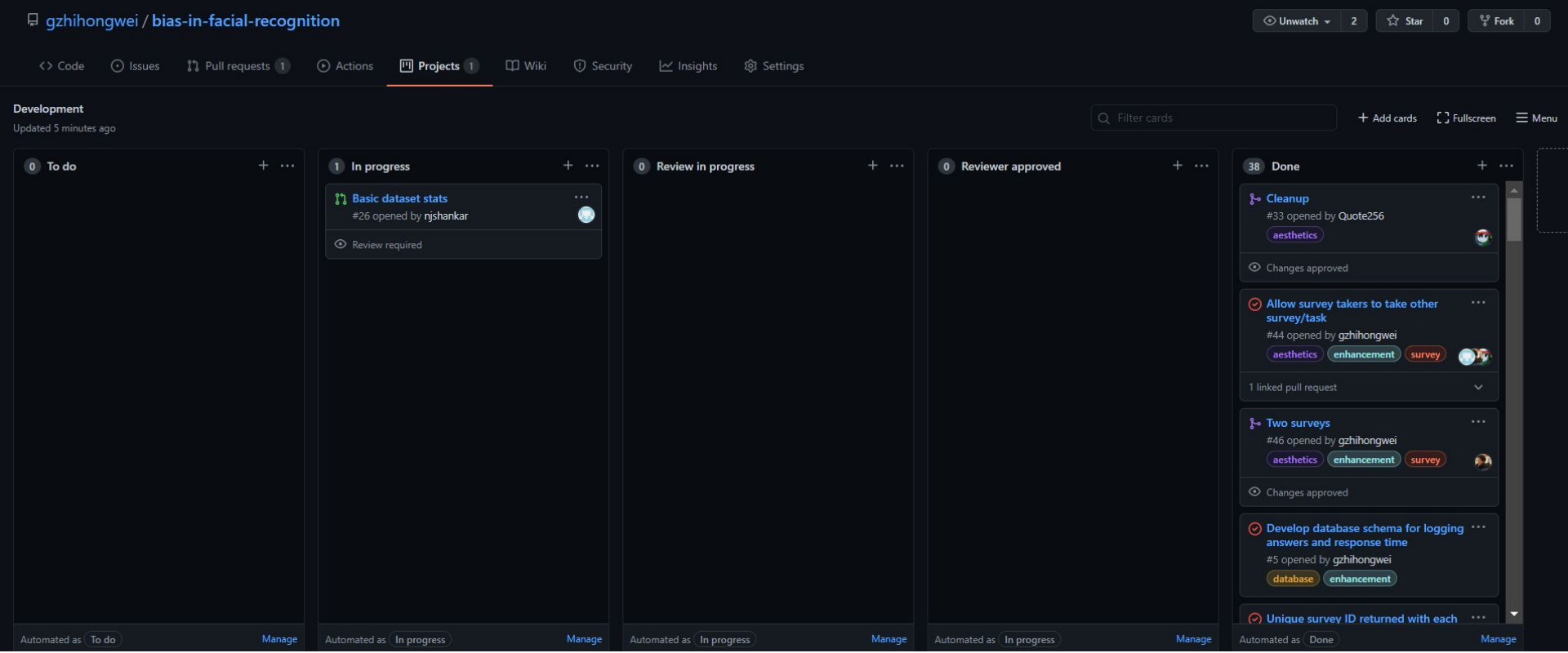
- Research shows that humans perform facial recognition tasks with bias
- People are better able to identify others belonging to their own racial group
  - Cross-Race Effect
- Similar effects can be seen across gender and age groups
- Raises the question: Is there a significant difference in the bias of facial recognition systems and human recognition?

Our Study

- The mission of our study is to quantify the extent of human bias in a way that is comparable to that of leading facial recognition algorithms.
  - Our verification test asks participants to verify whether two images are the same person
  - Our identification test asks participants to select a particular person's face from a lineup of 9 images.
- The same test will be performed on leading facial recognition technologies, and the results between humans and computer will be compared.

App

- Built a Flask web application with Bootstrap 5 to make the survey website
  - Flask is a Python web framework which allows rapid development of websites
  - Bootstrap 5 is a CSS library that defines many useful stylized classes
  - Uses vanilla JavaScript for dynamic elements
- Used the agile framework to develop the application (link to the repo here)
  - Agile is a development framework that focuses on self-defined tasks, a task board, and frequent communication with the stakeholders



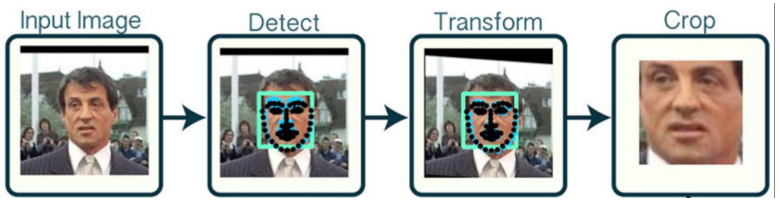
Kanban board used to track survey development progress

Datasets

- Use **Labelled Faces in the Wild (LFW)** as initial source of test question images.
- Supplemented set of LFW images with photos from **Celeb-A** dataset
- Both datasets are research standards, contain photos of celebrities.
- Chose LFW as **main** dataset due to manageable size and mostly accurate labelled identities.

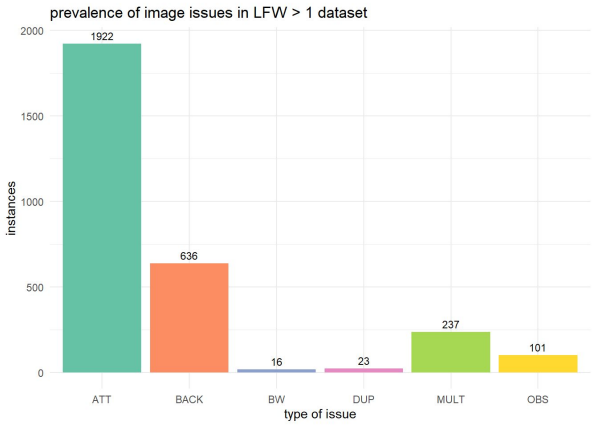
Datasets

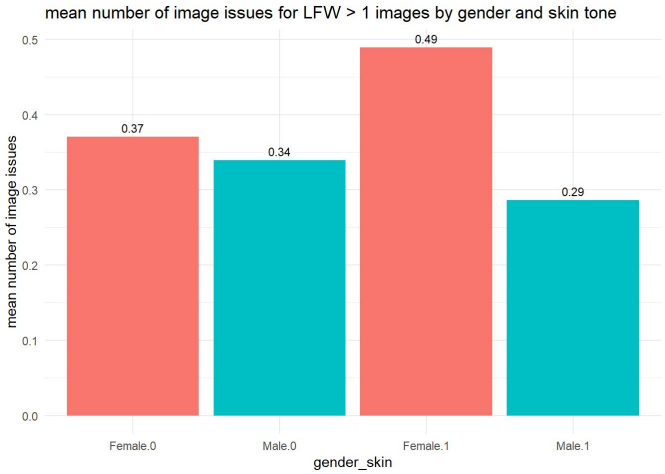
- Some **issues** with both LFW and Celeb-A
  - Low representation of certain groups (LFW particularly)
    - Comparatively lower percentage of women
    - Small range of ethnicities
    - Minimal to no children and babies.
  - Duplicate pairs of images.
  - Images with more than one person.
  - Images with the face being blocked.
  - Images clearly taken on the same occasion
    - particularly unsuitable for human testing



Datasets

- **Tagged** issues with images in LFW.
  - ATT/BACK/BW/DUP/MULT/OBS
- **Filtered** out images with issues for human test design.
  - Interesting (but preliminary finding): disparities in mean number of image issues depending on gender of image subject.





Datasets

- Since our study is concerned with sensitive attributes of images, we require labels for these attributes.
  - Some label datasets exist for LFW (such as MSU's **LFW+**), and Celeb-A comes with attribute labels, but these are incomplete and contain several clear mislabellings.
- We manually labelled all images for subject's **birthdate**, **country** of origin, **gender** presentation, and **skin tone**.
  - Only kept images for which these attributes could be exhaustively verified (i.e. through media articles or official biographies).
  - Classified skin tones by **Fitzpatrick scale**. Ultimately decided by a 2/3 majority vote.
  - Skin tone does not correlate perfectly with race, but more objective than race.

Datasets

- Final dataset to contain approximately **7500** images.
  - At least **100** pairs for each combination of gender presentation (male, female, other), and Fitzpatrick skin tone (1 to 6).
  - Strong coverage across ages within each gender-skin group too.
- **Beneficial Outcome: Cleaned-up** and **methodically-labelled** dataset
  - Could be useful for future facial recognition studies, in addition to this one.






Database Model

- Records survey responses
  - Unique user ids
  - Stores question metadata (images included, difficulty)
  - User identification information (age, skin tone, country)
- SQL database setup on UMIACS
  - Python API communicates between Flask and SQL
  - Flask submits form response to API
  - API validates data and inserts into database
- Next steps
  - Gather survey responses
  - Aggregate data to test hypothesis and produce statistics

LOOK! IT WORKS!

Anecdotal Evidence

- Identification is much more difficult than verification
- Some found caucasians harder to identify than people of other demographics
-

Design Decisions

- Survey format
  - Verification and Identification
  - Gallery of 9 (difficult but not overwhelming). Corresponds closely to police lineup.
- Number of questions
  - Internal trials for length
  - Try to shoot for equal number of each demographic bucket
  - Make questions as difficult as possible (same Fitzpatrick in gallery, same gender, similar age window)
- Collecting user information
  - Age, gender presentation, Fitzpatrick tone, country of origin.
  - Fitzpatrick images.

Ideas for Analysis/Future Plans

Recap - we are interested in comparing AI and human performance in a way that highlights weak areas for each.

- Once we get responses:
  - Compare AI vs. Human overall for each task
  - Look at target groups and see which have worst performance on them
  - Compare human test-taking groups to see if there is in-group/out-group bias.
- Share our dataset and welcome discussion

Thanks for Listening

Suggestions/Feedback/Comments?