In 2018, MIT found that facial recognition tools from large public companies had large discrepancies in face matching rates among sex and historical ethnic subgroups. Darker-skin female faces were the least reliably recognized among all sex and historical ethnic subgroups.

### Types of Bias

1. Sampling bias: a mismatch between the demographic makeup of a dataset and the population it was sampled from.
2. Historical bias: derived from previous generational biases that have to be manually corrected.
3. Representational bias: a bias within a population itself.
4. Application bias: a mismatch between the training dataset and the testing dataset.
5. Algorithmic bias: directly from the nature of the algorithm.

### Sampling Bias Tool

- Our tool counts the number of faces in each sex and historical ethnic group then uses a chi-square goodness of fit test for a multinomial PDF to ensure the sample is representative of the population it was sampled from (right lower).
- It analyzes a training dataset and outputs a bar graph showing the number of faces in each subgroup to display the bias (left).
- The bias score determines if a particular subgroup was oversampled or undersampled in the dataset, relative to the population.

### Application Bias Tool

- Our tool checks that the training and testing datasets are similar by computing a cumulative angular shift of the principal components of the training set relative to the testing set (lower).
- Individual test samples are projected onto the principal components of the training set and the bias score is the outlier probability based on analyzing the quantities (right).

### Algorithmic Bias Tool

- Our tool implements Equalized Odds metric [2]. It quantifies the cumulative bias in an AI model's output (may include biases induced from the dataset).
- A face recognition model satisfies this metric if all subgroups have equal true positive and false positive rates.
- The bias score is the maximum absolute difference between the marginal and conditional output densities.

### Historical Bias Tool

- Our tool assigns a BRISQUE image quality score to each photo and finds the average image quality for each sex and historical ethnic subgroup (lower).
- It analyzes a training dataset and outputs a pie chart showing the average image qualities for each subgroup to display the bias (right).
- The bias score is determined by calculating the largest difference between the average image qualities of any two subgroups.

### The Effects of Biased Data on an Algorithm

- This experiment analyzes the relationship between algorithmic bias and the bias in a training dataset.
- We purposely injected bias into a dataset by adding artifacts to degrade quality and re-sampling certain subgroups.
- We plot a graph of the algorithmic bias, as measured by Equalized Odds, against the injected bias levels.
- ResNet and VGG16 models are used for the experiments, along with a cosine similarity score to match the extracted facial features (below left).
- The results are intuitive, i.e the bias in the model's output positively correlated with the bias in the dataset (below right).

### Policy Recommendations

- When an AI decision will result in an action of consequence (such as an arrest) it should be reviewed by a human first to ensure the decision was made correctly.
- AI practitioners should train models with datasets that use self-reported labels so that models don't incorrectly infer identities and thus learn based off of incorrect attributes.
- Training datasets and AI models should exhibit bias under the limits set by our project to ensure that they only carry negligible bias toward any group.

### Future Work, References, and Acknowledgments

- Improving computation speed for analyzing large-scale datasets.
- Building tools to eliminate or mitigate bias after detecting it.
- Further experimentation to explore other ways that biased datasets can affect bias in an algorithm.