CS 4873: Computing, Society & Professionalism

Blair MacIntyre | Professor | School of Interactive Computing

Week 13: AI, Algorithms, and Bias

April 12, 2021

Slides adapted from Sauvik Das, Munmun de Choudhury, and Amy Bruckman

Term Papers

- Presentations during discussion sections next week
- Term papers due 4/26
  - Cannot use “late days” for the paper

What is FATE?

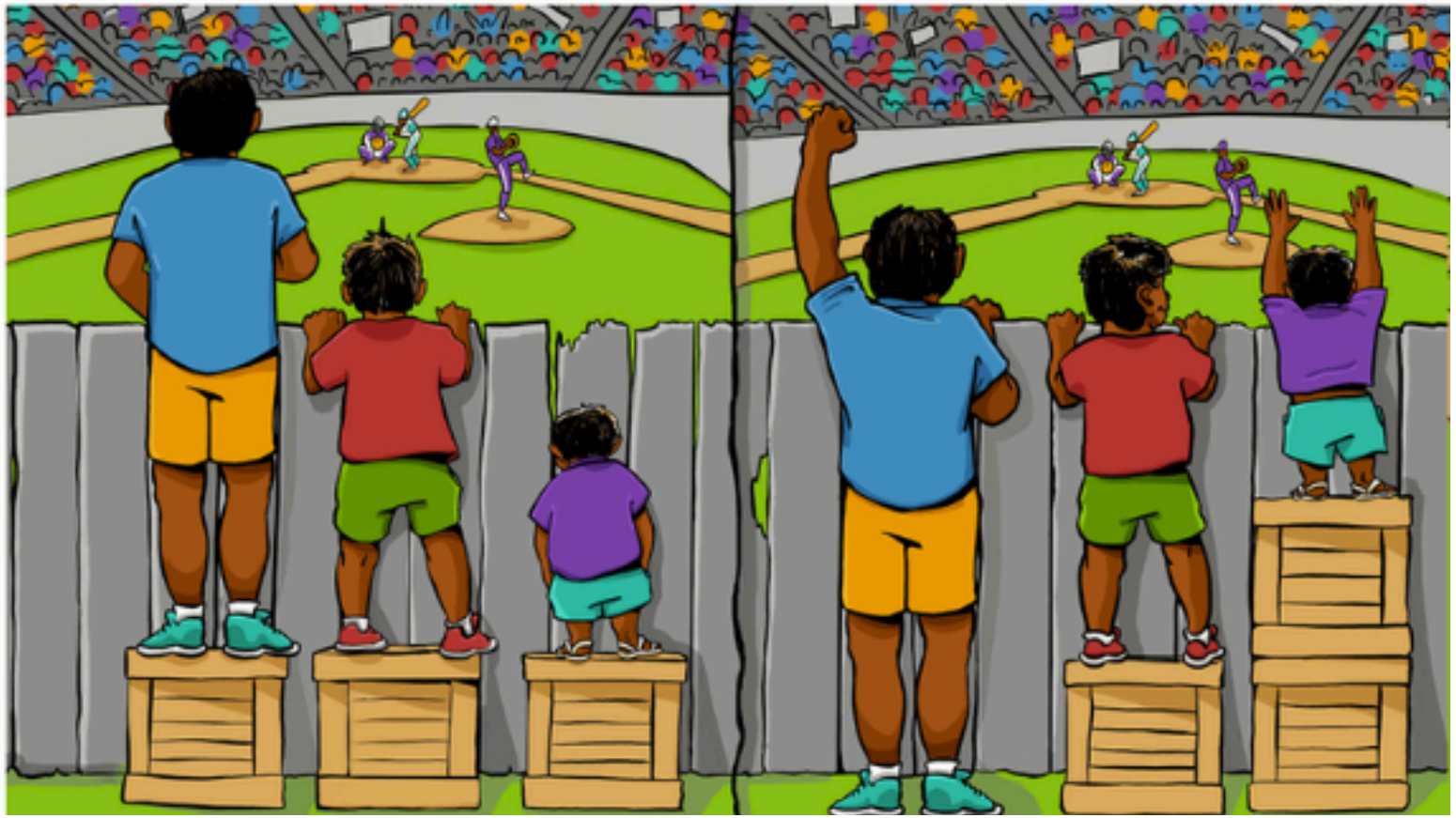
Ethical approach to AI:

- **F**airness
- **A**ccountability
- **T**ransparency
- **E**thics

Fairness and bias

- What is fair?

Which one of these allocations of crates would you consider to be fair?

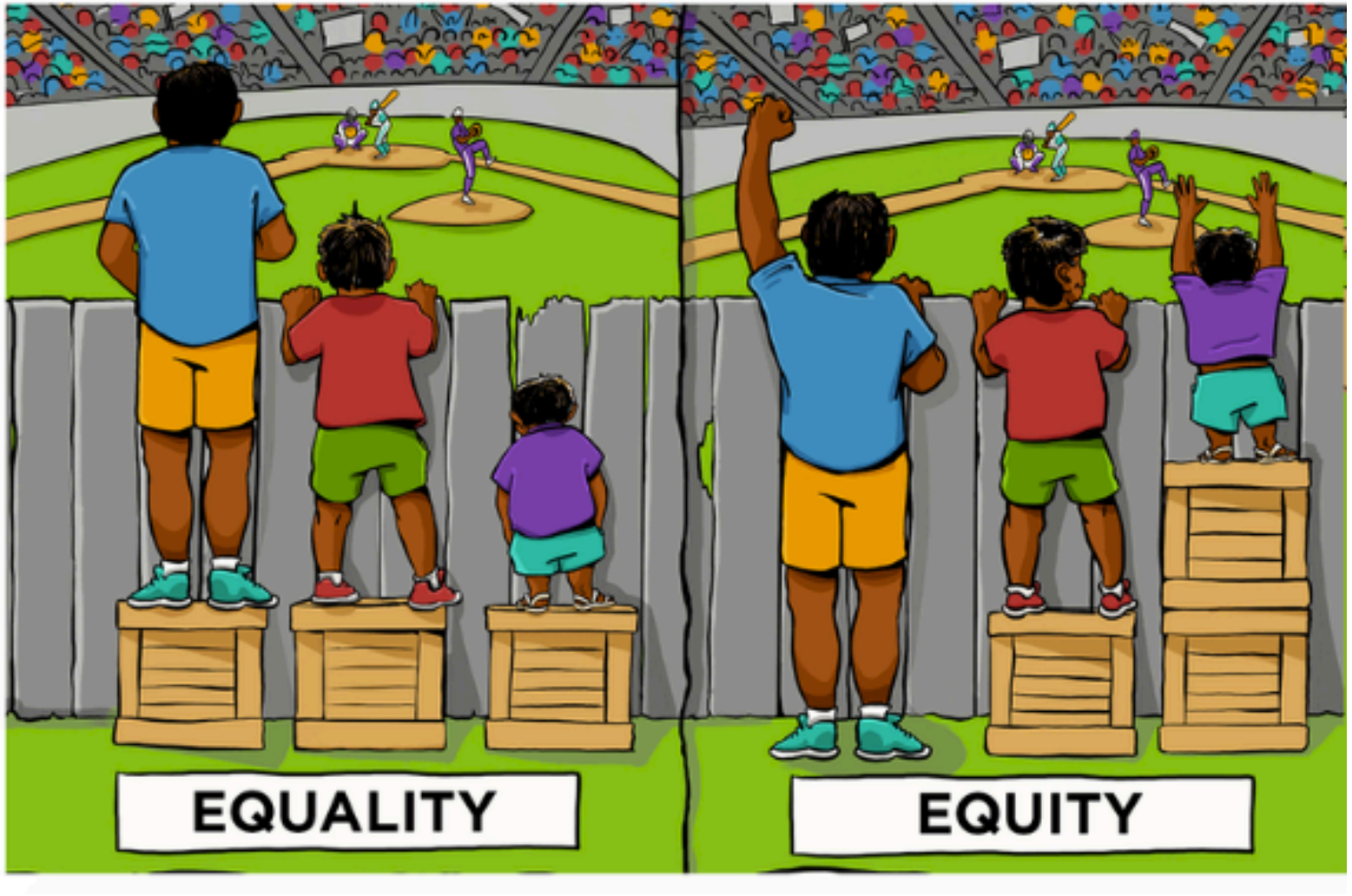


Is fairness the same thing as equality?

- Consider the following instances of “unequal” treatment that many consider ”fair”:
  - Children being given more slack than adults
  - Charging out-of-state students more than in-state for tuition
  - Taxing rich people more than poor people
  - Equal opportunity hiring / access
- Fairness is less about equality of treatment and more about equity of outcome

Equal but not fair

Unequal but fair

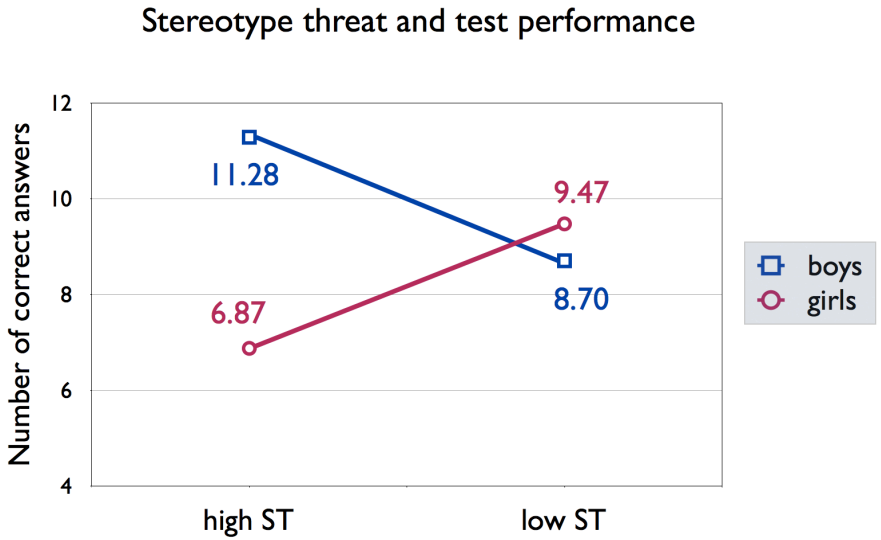


What is Bias?

- Inclination or prejudice for or against one person or group, especially in a way considered to be *unfair*
- Often tied to preconceived notions about a person's gender, race, age, or sexual orientation
- Can be both positive and negative
  - Negative bias: racism
  - Positive unfair bias: tall, good looking baseball players used to get drafted higher
    - Until the advent of the systematic use of statistics
    - See *Moneyballl by* Michael Lewis

Implicit Bias

- We may not be aware that we are biased
- Stereotype Threat
  - Example of experiment:
    - Introduce math test and describe it as hard
      - Men outperform women
    - Introduce math test & say no gender differences in performance
      - Equal performance by gender



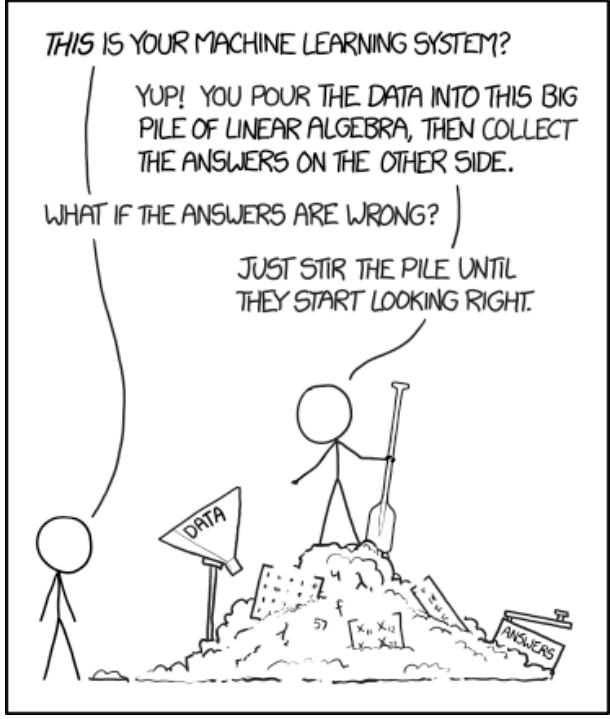
From https://en.wikipedia.org/wiki/Stereotype_threat

Implicit Association Test

- Sample items
  - Aunt, Son, Geology, Music
- Test 1:
  - Click e for woman or humanities
  - Click I for male or science
- Test 2:
  - Click e for woman or science
  - Click I for male or humanities
- Reaction time is slower for test 2, for most people

We Are All Somewhat Biased

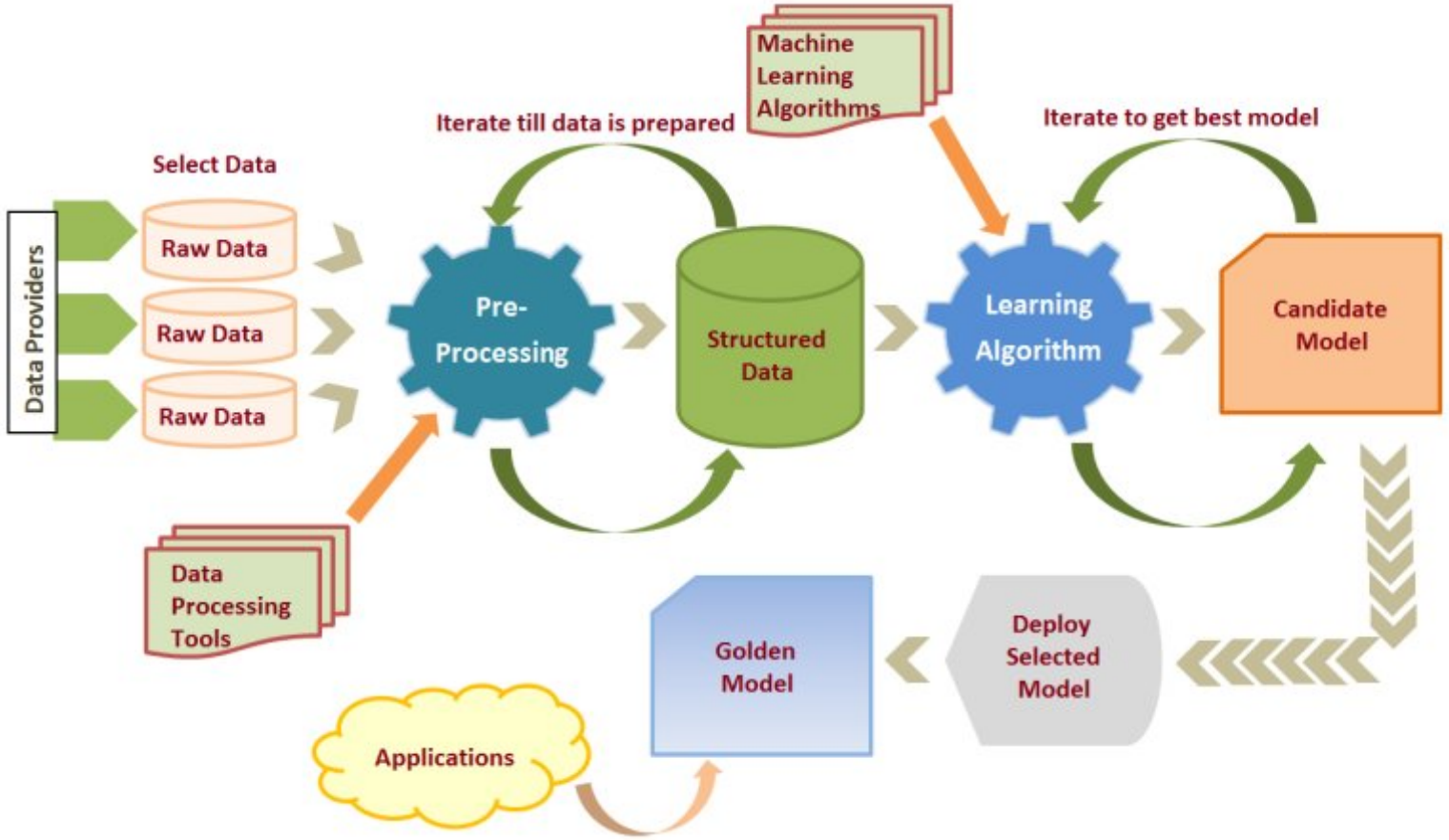
https://xkcd.com/1838/



ML Versus Symbolic AI

- Machine Learning (ML) takes a dataset and uses statistical methods
- Symbolic AI reasons about symbols that have real-world meaning
- Trend right now is towards ML approaches
- Eventually will need a synthesis

Broad ML Pipeline



https://www.kdnuggets.com/2018/12/essence-machine-learning.html

Two areas of concern: data and algorithms

Data inputs:

- Poorly selected (e.g., observe only car trips, not bicycle trips)
- Incomplete, incorrect, or outdated
- Selected with bias (e.g., smartphone users)
- Perpetuating and promoting historical biases (e.g., hiring people that "fit the culture")

Algorithmic processing:

- Poorly designed matching systems
- Personalization and recommendation services that narrow instead of expand user options
- Decision making systems that assume correlation implies causation
- Algorithms that do not compensate for datasets that disproportionately represent populations
- Output models that are hard to understand or explain hinder detection and mitigation of bias

Executive Office of the US President (May 2016): "Big Data: A Report on Algorithmic Systems,Opportunity,and Civil Rights"

Types of Bias in ML Datasets

- Reporting bias
  - Example: book reviews are strongly positive and negative
- Automation bias
  - People believe results from an algorithm more, even if it’s not merited

Selection Bias

- Data selected is not representative
- Coverage bias
  - Data selected in a non-representative way
  - Example: model trained on surveys with our company's product, not including competitors
- Non-response bias
  - People who bought the product are more likely to respond to the survey
- Sampling bias
  - Use first 200 responses to an email, instead of randomly selecting customers

Group Attribution Bias

- In-group bias
  - Bias in favor of people like you
  - Example: hiring a fellow GT grad
- Out-group homogeneity bias
  - Tendency to stereotype people not in your group

Types of Bias in ML Datasets, cont.

- Implicit bias
  - Assumptions based on your own assumptions that don’t hold more generally
  - Example: assuming head shake means “yes”
- Confirmation bias
  - Unconsciously affirm your own beliefs
- Experimenter’s bias
  - Keep refining what you’re building until it gives the answer you expected

ML and Bias

- Black boxed algorithms
  - It gave an answer, but why?
- Garbage in, garbage out
  - If the data is bad, the results will be bad

Open the Black Box: Explainable AI

- Active area of research
- ML models should be able to generate a human-understandable reason why they make a decision
  - One proposed approach: learn a rule-based expert system from model
  - Problem: would the “rules” make sense to humans?

Biased data

Biased AI

Bias in = Bias out

Biased humans

Biased algorithms

Search Engine Bias

- Ads for executive jobs are displayed less often to women
- Search for “doctors” shows mostly men
  - But 34% of doctors are women
  - Has improved since this was pointed out
- Search for “three white teenagers”→happy kids
- Search for “three black teenagers”→mug shots

[The ‘three black teenagers’ search shows it is society, not Google, that is racist | Search engines | The Guardian]

On the web: race and gender stereotypes reinforced

- Results for "CEO" in Google Images: 11% female, US 27% female CEOs
  - Also in Google Images, "doctors" are mostly male, "nurses" are mostly female
- Google search results for professional vs. unprofessional hairstyles for work

Image results: "Unprofessional hair for work"

Image results: "Professional hair for work"

M. Kay, C. Matuszek, S. Munson (2015): Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. CHI'15.

AI and the Justice System

- Parole and sentencing recommendations
- Predictive policing
  - Predict crimes
  - Predict offenders
  - Predict perpetrators identities
  - Predict victims
- Both tend to reinforce racist status quo
- Training data documents past racist practices

Judiciary use of COMPAS scores

COMPAS (Correctional Offender Management Profiling for Alternative Sanctions):
137-questions questionnaire and predictive model for "risk of recidivism"

Prediction accuracy of recidivism for blacks and whites is about 60%, but ...

- Blacks that did not reoffend
  were classified as high risk twice as much as whites that did not reoffend
- Whites who did reoffend
  were classified as low risk twice as much as blacks who did reoffend

Pro Publica, May 2016. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Predict Criminality from Face

- Quote from researchers who created the classifier to predict criminality from face:
  - “Unlike a human examiner/judge, a computer vision algorithm or classifier has absolutely no subjective baggages…The automated inference on criminality eliminates the variable of meta-accuracy (the competence of the human judge/examiner) all together.”

(a) Three samples in criminal ID photo set S_c.

(b) Three samples in non-criminal ID photo set S_n

Figure 1. Sample ID photos in our data set.

Bias and Robotics

- Military robots can make life or death decisions
  - Failures of computer vision on dark skin make fatal errors more likely
- Driverless cars
- Medical robots
  - Medicine requires value judgments
    - Example: triage

Deep neural networks are more accurate than humans at detecting sexual orientation from facial images

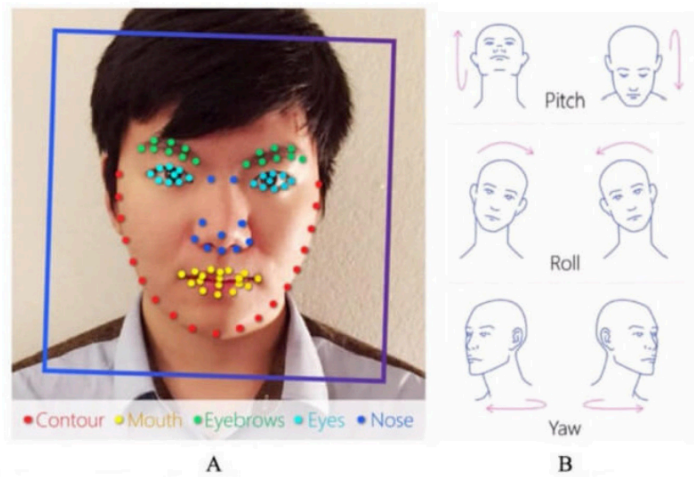
- Authors used deep neural networks to extract features from 35,326 facial images.
  - Images scraped from public profiles posted on a U.S. dating website
- These features were entered into a logistic regression aimed at classifying sexual orientation.
- Given a single facial image, a classifier could correctly distinguish between gay and heterosexual men in 81% of cases, and in 74% of cases for women.
- The authors claimed that their findings therefore provided “strong support” for the idea that sexual orientation stems from hormone exposure in the womb

SCIENCE

The Study Claiming AI Can Tell If You're Gay or Straight Is Now Under Ethical Review

By Lisa Ryan @lisarya

SEPTEMBER 12, 2017
6:21 PM



An image from the study. Photo: Journal of Personality and Social Psychology/Stanford University

A recent Stanford University study published in the *Journal of Personality and Social Psychology* claimed artificial intelligence can figure out if a person is gay or straight by analyzing pictures of their faces. However, the Outline reports the study was met with "immediate backlash" from the AI community, academics, and LGBTQ advocates alike — and the paper is now under ethical review.

Complexities of debate

- The researchers have argued that they were simply demonstrating that AI *could* be used to this, and that it could have dangerous applications
- The results are the results. Critiques should be evidence-based not rhetorical.

Physiognomy: the (discredited) notion that personality traits can be revealed by measuring the size and shape of a person's eyes, nose and face.

AI-facilitated physiognomy: using AI to predict personality traits, sexual orientation, criminality etc.

Based on Assumption: This correlation really does exist; humans too imprecise and biased, but AI can overcome that.

Questions

- Was this use of ML to predict sexual orientation unfair?
  - Why or why not?

- What kind of biases can this sexual orientation detector that uses facial images introduce in platforms that rely on profiling users?

The ethical challenges

- Algorithmic bias is shaping up to be a major societal issue at a critical moment in the evolution of machine learning and AI.
- If the bias lurking inside the algorithms that make ever-more-important decisions goes unrecognized and unchecked, it could have serious negative consequences, especially for marginalized communities and minorities.