weizenbaum institut

Technische Universität Berlin

WEIZENBAUM INSTITUTE

FOR THE NETWORKED SOCIETY

Gender and Racial bias in AI-based systems

Gunay Kazimzade
Doctoral Researcher
Technical University of Berlin

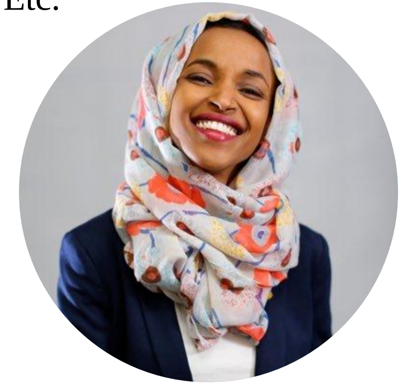
ACM WomENcourage 2019
Rome, Italy

JOINT PROJECT

weizenbaum institut

Imagine that you are preparing to train a classifier with the set of images you label manually(3-5 labels per image).
Labels should be nouns and/or adjectives including but not limited to:

- Possible occupation
- emotional state
- visual components
- Etc.



Labels:

- happy
- young
- smiley
- covered
- dressed up

Labels:

- singer
- politician
- proud
- curly
- rich

Labels:

- serious
- senior
- sad
- worried
- black

Labels:

- white
- male
- angry
- old
- glasses

weizenbaum institut

Imagine that you are preparing to train a classifier with the set of images you label manually(3-5 labels per image).
Labels should be nouns and/or adjectives including but not limited to:

- Possible occupation
- emotional state
- visual components
- Etc.

Labels:

- politician
- black
- muslim
- smiley
- self confident

Labels:

- experience
- proud
- fighter
- politician
- tough

Labels:

- smiley
- friendly
- bald
- positive
- black

Labels:

- tiny
- patriotic
- white
- elegant
- confused

weizenbaum institut

Inclination or prejudice for or against one person or group, especially in a way considered to be unfair.

Amazon Reportedly Killed an AI Recruitment System Because It ...
Fortune - 10 Oct 2018
Amazon Reportedly Killed an AI Recruitment System Because It Couldn't Stop the Tool from Discriminating Against Women ...
Amazon used AI to promote diversity. Too bad it's plagued with gender ...
In-Depth - Mashable - 10 Oct 2018
View all

Amazon scraps 'sexist AI' recruitment tool
The Independent - 11 Oct 2018
Amazon has scrapped a "sexist" tool that used artificial intelligence to decide the best candidates to hire for jobs. Members of the team working ...
Amazon Shuts Down Secret AI Recruiting Tool That Taught Itself to be ...
Interesting Engineering - 12 Oct 2018
View all

Toronto police have been using facial recognition technology for more than a y

Facial recognition is coming to US schools, starting in New York

The first school district in the US to pi

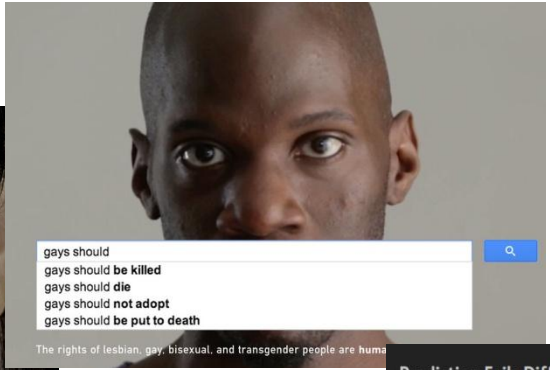
Brooklyn Landlord Wants To Install Facial Recognition Tec Rent-Stabilized Complex

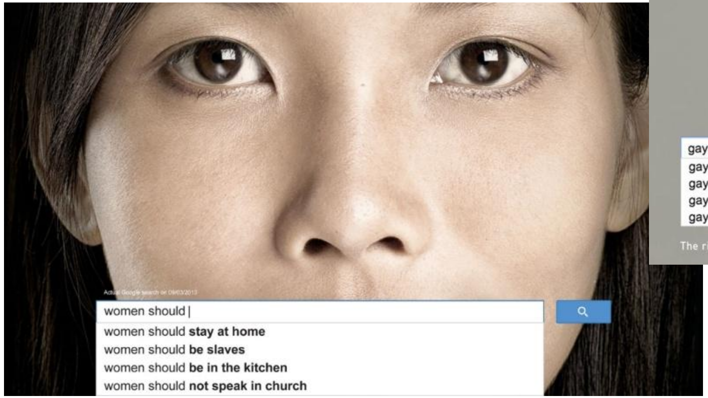
BY ELIZABETH KIM IN NEWS ON MAR 25, 2019 11:19 AM

Levi @xlevix10 1m
@TayandYou ARE YOU A RACIST?!

in reply to @xlevix10
TayTweets @TayandYou
@xlevix10 because ur mexican
7:01 PM - 23 Mar 16

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%





Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

BIASED AI vs.
BIASED SOCIETY

« AI can **amplify** discrimination and biases, such as **gender or racial discrimination**, because those are present in the data the technology is trained on, reflecting people's behaviour. »

Yoshua Bengio, 2019

Face recognition researcher fights Amazon over biased AI

There's software COMPAS used across the US to help judges in courtrooms forecast which criminals are most likely to reoffend. And **it's biased against blacks**."

Source : https://www.technologyreview.com/s/607955/inspecting-algorithms-for-bias/

Should We Let Data Speak for Itself?

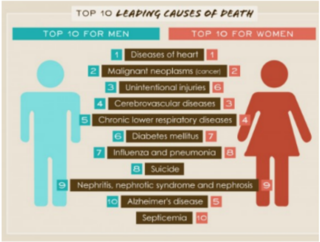
Data Quality Issues
Data Bias

Bias must be considered relative to task

Gender in loan application

Gender discrimination is illegal

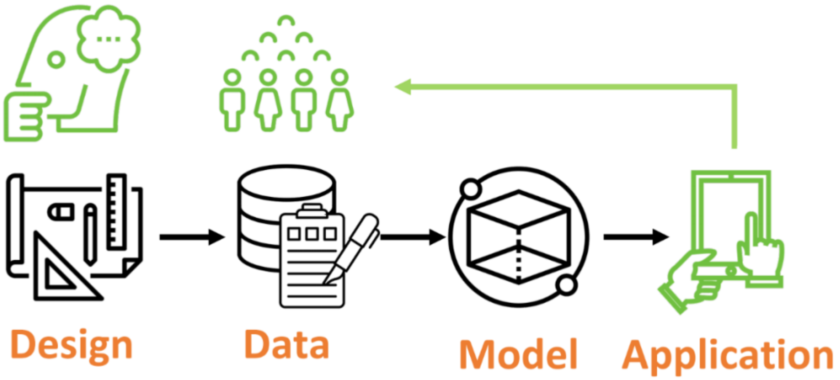
Gender in medical diagnosis



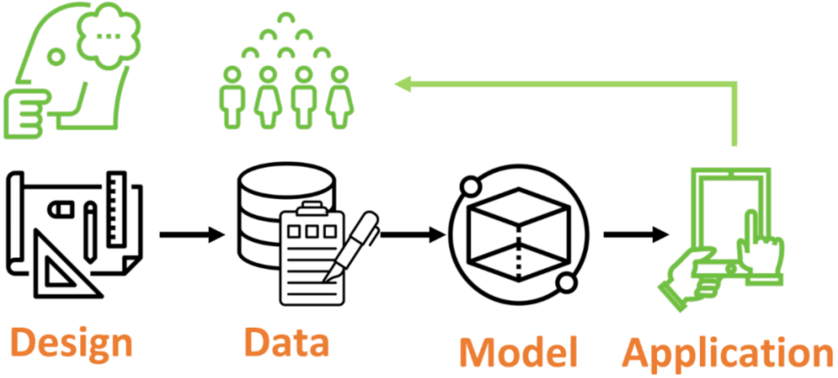
Gender-specific medical diagnosis is desirable 15

Where does the bias come from? Traditional Approach

DETECTING
MEASURING
MITIGATING



Where does the bias come from?



Bias can come in any step along the data analysis pipeline

Data Source

- **Functional:** biases due to platform affordances and algorithms
- **Normative:** biases due to community norms
- **External:** biases due to phenomena outside social platforms
- **Non-individuals:** e.g., organizations, automated agents

Data Collection

- **Acquisition:** biases due to, e.g., API limits
- **Querying:** biases due to, e.g., query formulation
- **Filtering:** biases due to removal of data “deemed” irrelevant

Data Processing

- **Cleaning:** biases due to, e.g., default values
- **Enrichment:** biases from manual or automated annotations
- **Aggregation:** e.g., grouping, organizing, or structuring data

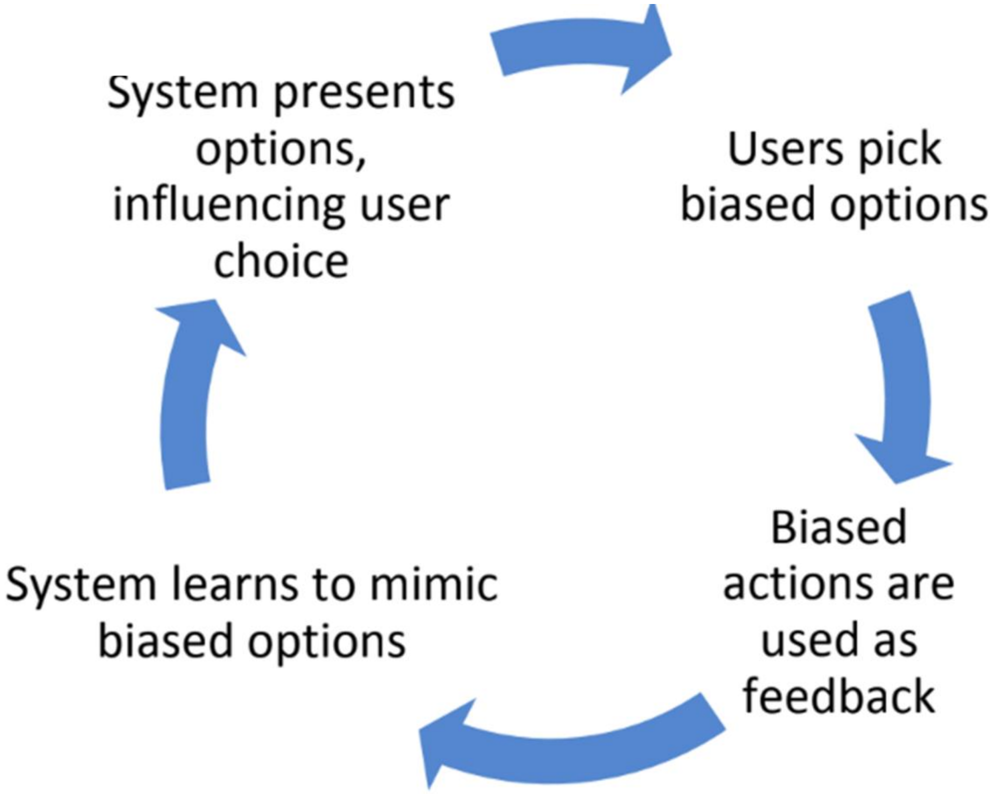
Data Analysis

- **Qualitative Analyses:** lack generalizability, interpret. biases
- **Descriptive Statistics:** confounding bias, obfuscated measurements
- **Prediction & Inferences:** data representation, perform. variations
- **Observational studies:** peer effects, select. bias, ignorability

Evaluation

- **Metrics:** e.g., reliability, lack of domain insights
- **Interpretation:** e.g., contextual validity, generalizability
- **Disclaimers:** e.g., lack of negative results and reproducibility

weizenbaum institut



What we have done so far?

- Categorization of biases and their representation in social datasets
- Data Labelling process analysis(ACM AI and Ethics) – Quality criteria from bias perspective

What we plan to do further?

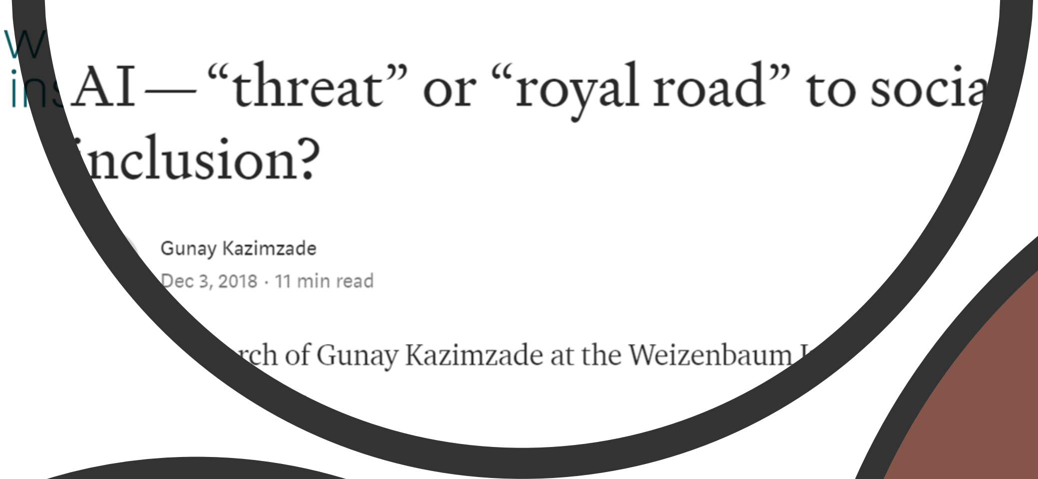
- Mathematical representation of this mapping(human biases to AI biases)
- Algorithmic representation
- Detecting bias in visual models(Framework and the system)

Encoded Stereotypes

Stock, Pierre and Moustapha Cissé. "ConvNets and ImageNet Beyond Accuracy: Understanding Mistakes and Uncovering Biases." ECCV (2018).

Approach

- Consider team composition for diversity of thought, background and experiences
- Understand the task, stakeholders, and potential for errors and harm
- Check data sets: Consider data provenance. What is the data intended to represent?
- Verify through qualitative, experimental, survey and other methods
- Check models and validate results: Why is the model making decision?
- What mechanisms would explain results? Is supporting evidence consistent?
- Twyman's law: The more unusual the result, more likely it's an error
- Post-Deployment: Ensure optimization and guardrail metrics consistent w/responsible practices and avoid harms.
- Continual monitoring, including customer feedback
- Have a plan to identify and respond to failures and harms as they occur



More in my TEDx speech

and Medium article

weizenbaum institut

JOSEPH WEIZENBAUM

“A society that engages in a technique needs a strong inner force in order not to be seduced by the goals, not to become too greedy.

- Joseph Weizenbaum

CONTACT

Weizenbaum Institute for the Networked Society
Hardenbergstr. 32, 10623 Berlin
www.weizenbaum-institut.de

Gunay Kazimzade
Gunay.kazimzade@tu-berlin.de