

AI BIAS: A SURVEY ON BIAS AND FAIRNESS IN MACHINE LEARNING

EIRIK BERGE

MARTIN LUDVIGSEN

CAMILLA ELVEBAKKEN

MA8701, NTNU

01.03.2021



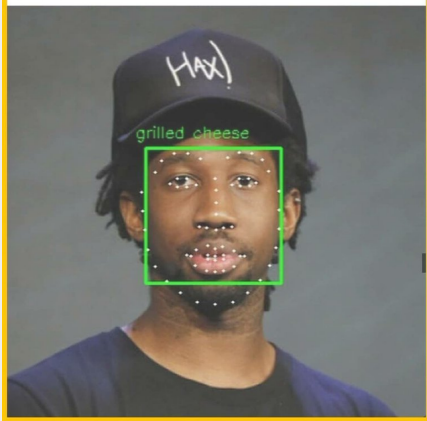
INTRODUCTION

INTRODUCTION

- **Is AI harmless?**
- Machines are superior to decision-making, but also vulnerable til biases
- Increasingly being commercialized and used in sensitive areas
- Must ensure that decisions do not reflect discriminatory behavior

Humanity: AI is going to take over the world soon

AI:



INTRODUCTION

Examples:

- COMPAS
- Amazon's hiring algorithm
- Health care-risk prediction algorithm
- Job advertisements

An **unfair** algorithm is one that is skewed towards a group or individual.

BIAS

BIAS

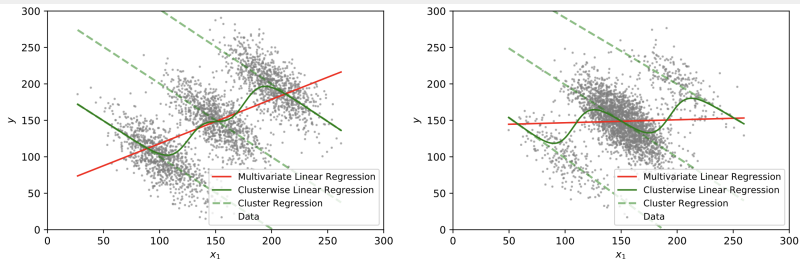


Fig. 1. Illustration of biases in data. Red line shows the regression (MLR) for the entire population, while dashed green lines are regressions for each subgroup, and the solid green line is the unbiased regression. (a) When all subgroups are of equal size, then MLR shows a positive relationship between the outcome and the independent variable. (b) Regression shows almost no relationship in less balanced data. The relationships between variables within each subgroup, however, remain the same. (Credit: Nazanin Alipourfard)

BIAS

Non-homogenous data → potentially biased data.

Examples of bias:

- **Simpson's paradox:** Previous figure.
- **Representation bias:** Sample some groups more than others.
- **Evaluation bias:** Inappropriate use of test benchmarks.
- **Observer bias:** Projecting own expectations onto research.

Problem: All parts of ML/AI research can be biased and we end up with a positive feedback loop! Bias is amplified!

BIAS

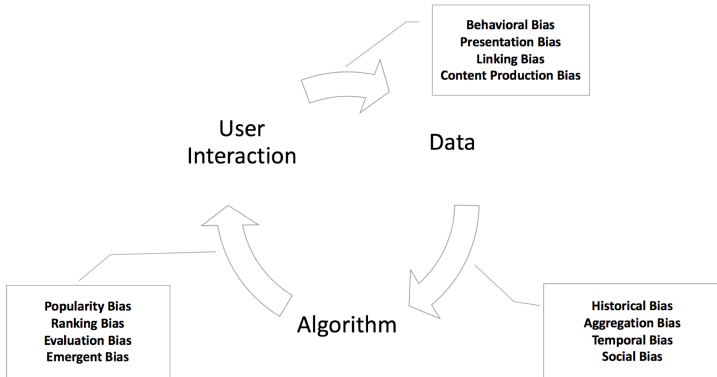


Fig. 2. Bias definitions in the data, algorithm, and user interaction feedback loop are placed on their most appropriate arrows.

DISCRIMINATION AND FAIRNESS

TYPES OF DISCRIMINATION

TYPES OF DISCRIMINATION

- **Direct Discrimination:** Protected attributes of individuals are explicitly used.

TYPES OF DISCRIMINATION

- **Direct Discrimination:** Protected attributes of individuals are explicitly used.
 - ▶ *Protected* here broadly means something that should not be used.

TYPES OF DISCRIMINATION

- **Direct Discrimination:** Protected attributes of individuals are explicitly used.
 - ▶ *Protected* here broadly means something that should not be used.
 - ▶ An example is using gender in an algorithm to determine legal punishment.

TYPES OF DISCRIMINATION

- **Direct Discrimination:** Protected attributes of individuals are explicitly used.
 - ▶ *Protected* here broadly means something that should not be used.
 - ▶ An example is using gender in an algorithm to determine legal punishment.

- **Indirect Discrimination:** Protected attributes of individuals are implicitly used through other variables.

TYPES OF DISCRIMINATION

- **Direct Discrimination:** Protected attributes of individuals are explicitly used.
 - ▶ *Protected* here broadly means something that should not be used.
 - ▶ An example is using gender in an algorithm to determine legal punishment.
- **Indirect Discrimination:** Protected attributes of individuals are implicitly used through other variables.
 - ▶ An example is zip codes or area of education to determine loan applications.

WHAT IS FAIRNESS?

WHAT IS FAIRNESS?

Fairness

Absence of any prejudice or favoritism towards an individual or a group based on their inherent or acquired characteristics is called **fairness**.

WHAT IS FAIRNESS?

Fairness

Absence of any prejudice or favoritism towards an individual or a group based on their inherent or acquired characteristics is called **fairness**.

There are many (empirical) ways of measuring fairness!

THREE WAYS TO DETERMINE FAIRNESS

THREE WAYS TO DETERMINE FAIRNESS

Fairness Through Unawareness - Easy

Protected attributes are not used explicitly.

THREE WAYS TO DETERMINE FAIRNESS

Fairness Through Unawareness - Easy

Protected attributes are not used explicitly.

Equal Opportunity Fairness - Medium

If \hat{Y} is a binary predictor and A is a protected attribute, then

$$P(\hat{Y} = 1 | A = 0, Y = 1) = P(\hat{Y} = 1 | A = 1, Y = 1).$$

THREE WAYS TO DETERMINE FAIRNESS

Fairness Through Unawareness - Easy

Protected attributes are not used explicitly.

Equal Opportunity Fairness - Medium

If \hat{Y} is a binary predictor and A is a protected attribute, then

$$P(\hat{Y} = 1 | A = 0, Y = 1) = P(\hat{Y} = 1 | A = 1, Y = 1).$$

Demographic Parity Fairness - Hard

If \hat{Y} is a binary predictor and A is a protected attribute, then

$$P(\hat{Y} | A = 0) = P(\hat{Y} | A = 1).$$

WHAT HAS BEEN DONE SO FAR?

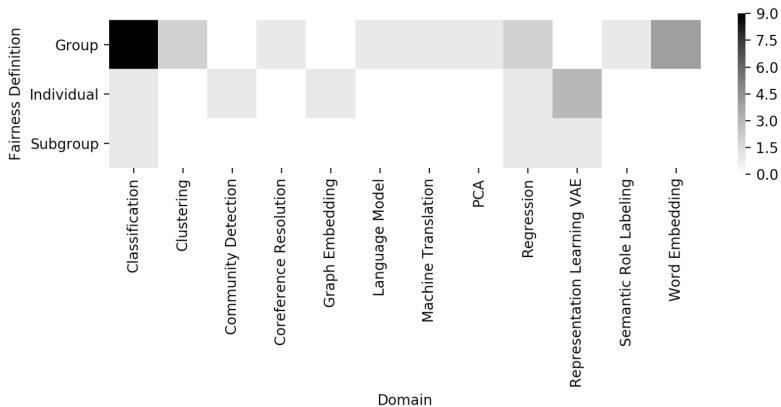


Fig. 7. Heatmap depicting distribution of previous work in fairness, grouped by domain and fairness definition.

IMPLEMENTING FAIRNESS IN AI

GENERAL APPROACH

- Implementation usually very **domain specific**
- Two sources of bias: data and algorithm
- Generally, three ways to deal with it

GENERAL APPROACH

Pre-processing

Transform the input data prior to training to remove bias.

Ex: re-weighting.

GENERAL APPROACH

Pre-processing

Transform the input data prior to training to remove bias.

Ex: re-weighting.

In-processing

Modify algorithm itself, to remove bias during training.

Ex: change objective function, add constraints

GENERAL APPROACH

Pre-processing

Transform the input data prior to training to remove bias.
Ex: re-weighting.

In-processing

Modify algorithm itself, to remove bias during training.
Ex: change objective function, add constraints

Post-processing

Reassign labels from black-box models after training to make outcome fairer.

GENERAL APPROACH

Algorithm	Reference	Pre-Processing	In-Processing	Post-Processing
Community detection	[85]	✓		
Word embedding	[22]	✓		
Optimized pre-processing	[26]	✓		
Data pre-processing	[63]	✓		
Classification	[131]		✓	
Regression	[13]		✓	
Classification	[65]		✓	
Classification	[128]		✓	
Adversarial learning	[76]		✓	
Classification	[55]			✓
Word embedding	[19]			✓
Classification	[103]			✓

Table 3. Algorithms categorized into their appropriate groups based on being pre-processing, in-processing, or post-processing.

GENERAL APPROACH

What can we do?

- Include fairness-penalization
- Include users from sensitive groups
- Avoid sensitive or protected attributes
- Good supporting documents for datasets:
 - ▶ creation method
 - ▶ characteristics
 - ▶ motivations
 - ▶ skews

GENERAL APPROACH

Tools:

- Automatic detection of Simpson's paradoxes in data
- Aequitas
- AI Fairness 360
 - ▶ <https://aif360.mybluemix.net/>

FAIR REGRESSION

Assume dataset $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ separated into two groups S_1 and S_2 based on sensitive data.

Add fairness penalization

$$\hat{\beta} = \arg \min_{\beta} \underbrace{l(\beta, S)}_{\text{standard loss}} + \underbrace{\lambda f(\beta, S)}_{\text{fairness penalization}}. \quad (1)$$

For regression, one example fairness penalization [1] is

$$f(\beta, S) = \frac{1}{|S_1||S_2|} \sum_{\substack{(\mathbf{x}_i, y_i) \in S_1, \\ (\mathbf{x}_j, y_j) \in S_2}} e^{-(y_i - y_j)} (\mathbf{x}_i^T \beta - \mathbf{x}_j^T \beta)^2 \quad (2)$$

- Split models $\beta_{S_1} \neq \beta_{S_2}$ or unified model $\beta_{S_1} = \beta_{S_2}$?
- Explicitly use sensitive data in training?

FAIR REGRESSION

- $\lambda = 0$: accurate, but potentially unfair model.
- $\lambda \rightarrow \infty$ completely fair, but constant (useless?) model.
- Accuracy-Fairness tradeoff.
- Different datasets have different sensitivity to imposed fairness.
- Authors define metric that measures the "price of fairness" that can be used to investigate sensitivity to imposed fairness.
- **Open question:** What is the "optimal amount" of fairness?

CHALLENGES AND FUTURE DEVELOPMENT

CHALLENGES AND FUTURE





Challenges:

- Unified definition of fairness
- (Automatic) unfairness detection
- Equality vs. equity

OUR RESPONSIBILITY

- We cannot expect end-users (doctors, judges, police, bank managers) without training in ML/AI/statistics will ensure that ML does not cause harm or bias towards particular groups.
- Data collecting/aggregation should always be of high quality.
- Researchers (like you!) are left with the responsibility to ensure that ML models avoid bias against particular groups.
- ...which motivates more research on bias and fairness!

REFERENCES

-  RICHARD BERK, HODA HEIDARI, SHAHIN JABBARI, MATTHEW JOSEPH, MICHAEL KEARNS, JAMIE MORGENSTERN, SETH NEEL, AND AARON ROTH.
A CONVEX FRAMEWORK FOR FAIR REGRESSION, 2017.
-  **ATAKAN KANTARCI.**
BIAS IN AI: WHAT IT IS, TYPES & EXAMPLES, HOW & TOOLS TO FIX IT.
<https://research.aimultiple.com/ai-bias/>, 2021.
[Online; accessed 23-February-2021].
-  CHRISTINE MAROTI.
GENDER BIAS IN AI: BUILDING FAIRER ALGORITHMS.
<https://unbabel.com/blog/gender-bias-artificial-intelligence/>, 2019.
[Online; accessed 23-February-2021].
-  NINAREH MEHRABI, FRED MORSTATTER, NRIPSUTA SAXENA, KRISTINA LERMAN, AND ARAM GALSTYAN.
A SURVEY ON BIAS AND FAIRNESS IN MACHINE LEARNING, 2019.

QUESTIONS?