A Deep Dive into Fairness in Machine Learning using Fairlearn

Hilde Weerts
Artificial Intelligence Engineer
Eindhoven University of Technology

Agenda

- Algorithmic Fairness
- Introduction to Tutorial

Introduction
Algorithmic Fairness

Algorithmic Fairness

The idea that algorithmic systems should behave or treat people without unjust or prejudicial treatment on the grounds of sensitive characteristics.

Hiring

RETAIL OCTOBER 11, 2018 / 1:04 AM / UPDATED 2 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin 8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's AMZN.O machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

Source: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Fraud Detection

XENOPHOBIC MACHINES

DISCRIMINATION THROUGH UNREGULATED USE OF ALGORITHMS IN THE DUTCH CHILDCARE BENEFITS SCANDAL

AMNESTY INTERNATIONAL

Source: https://www.amnesty.nl/content/uploads/2021/10/20211014_FINAL_Xenophobic-Machines.pdf?x42580

Translation

Turkish - detected	English
o bir aşçı	she is a cook
o bir mühendis	he is an engineer
o bir doktor	he is a doctor
o bir hemşire	she is a nurse
o bir temizlikçi	he is a cleaner
o bir polis	He-she is a police
o bir asker	he is a soldier
o bir öğretmen	She's a teacher
o bir sekreter	he is a secretary

Source: https://qz.com/1141122/google-translates-gender-bias-pairs-he-with-hardworking-and-she-with-lazy-and-other-examples/

Types of Harm

Majority of fairness research focuses on these two harms

Allocation

The system extends or witholds opportunities, resources, or information.

Quality-of-Service

The system does not work equally well for all groups.

Representation

The development/usage of the system overrepresents or underrepresents certain groups.

Stereotyping

The system reinforces stereotypes.

Most prevalent in unstructured data

Denigration

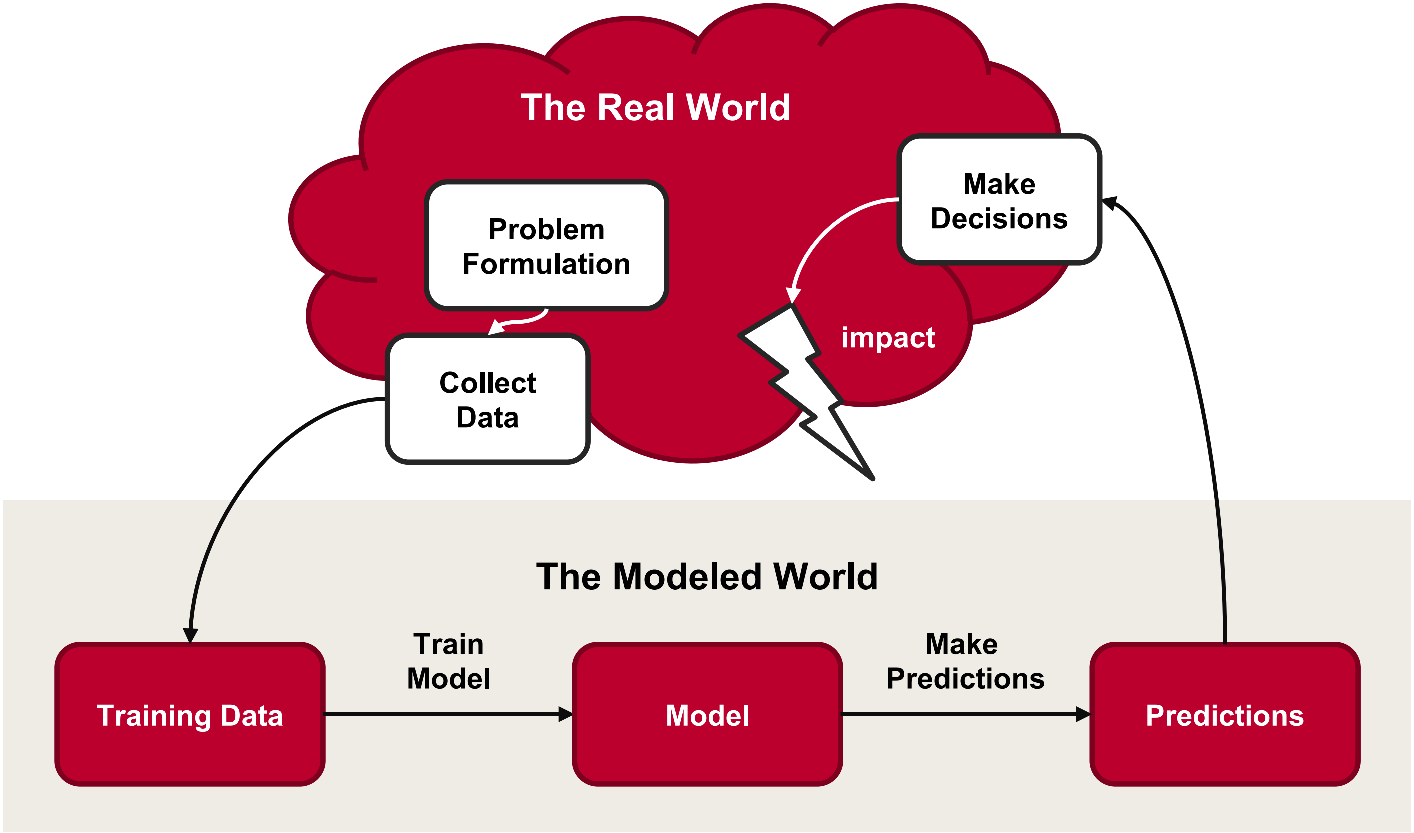
The system is actively derogatory or offensive.

Procedural

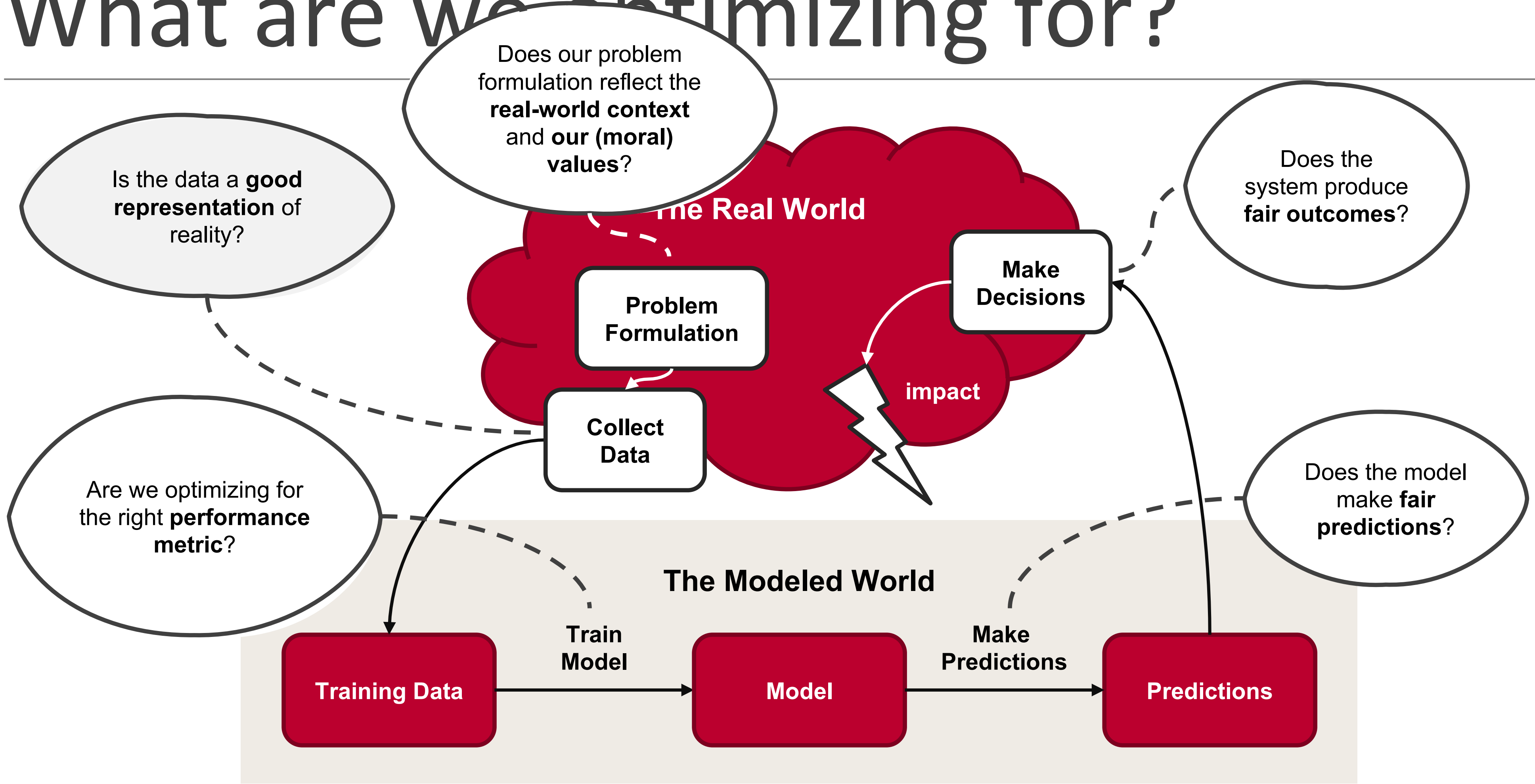
The system makes decisions in a way that violates social norms.

Closely related to interpretable machine learning

What are we optimizing for?

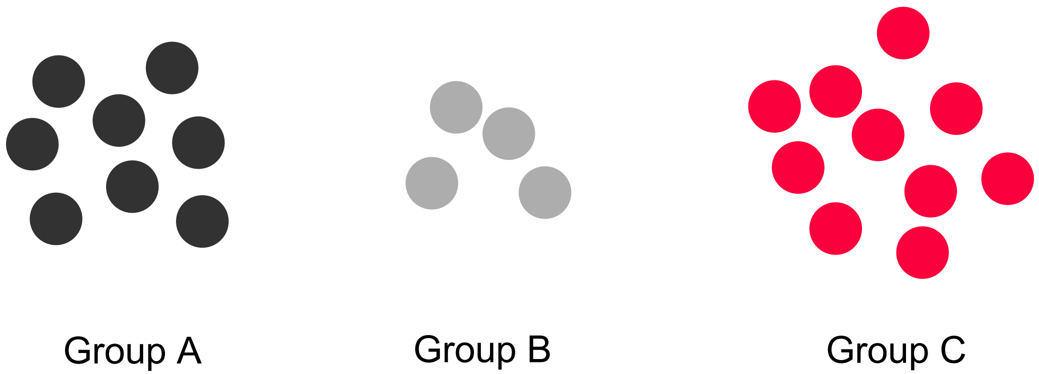


What are we optimizing for?



Group Fairness

Notions of fairness that **compare** the behavior or performance of a model across **sensitive groups**.



A **group fairness metric** measures the extent to which a particular **group statistic** differs across groups (e.g., *maximum difference between groups*).

Demographic Parity

Group 1
SR = 3/5 = 0.6

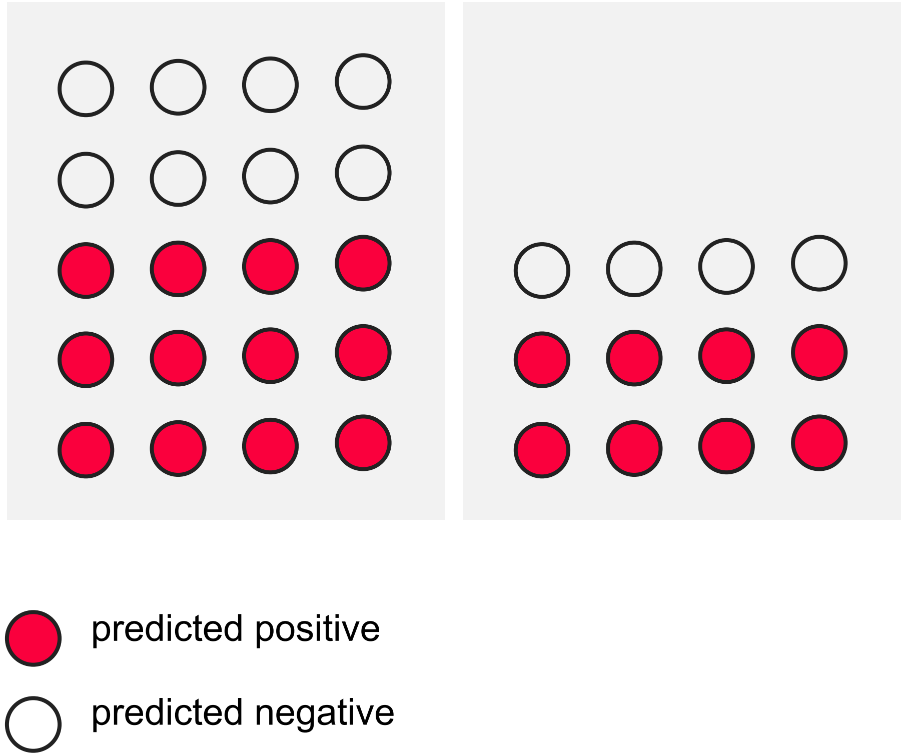
Group 2
SR = 2/3 = 0.67

Fairness criterion: **equal selection rates**

$$P(\hat{Y} = y \mid A = a) = P(\hat{Y} = y \mid A = a')$$

Measures: **allocation harm**

Main assumption: the target variable is not a good representation of reality and/or what we want reality to look like.



Equalized Odds

Group 1
FPR = 3 / (3+7) = 0.3
FNR = 1 / (1+9) = 0.1

Group 2
FPR = 3 / (3+4) = 0.43
FNR = 0 / (0+5) = 0

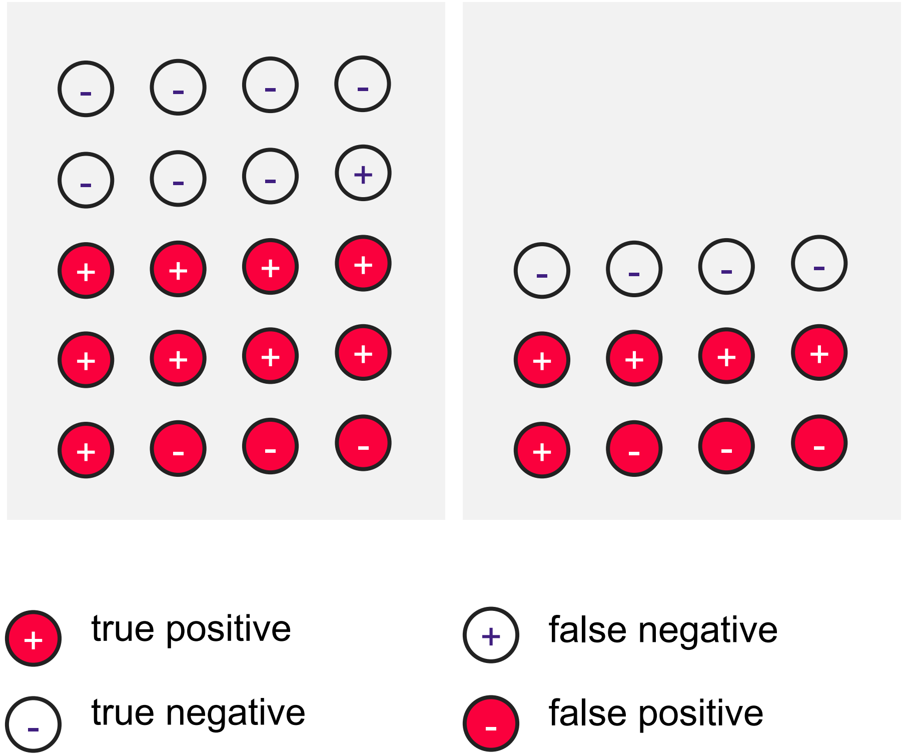
Fairness criterion: **equal FPR and FNR**

$$P(\hat{Y} = y | A = a, Y = y) = P(\hat{Y} = y | A = a', Y = y)$$

Measures: **quality-of-service harm** or **allocation harm**

Main assumptions:

- the target variable is a good representation of reality
- the data distribution may differ across groups, which could cause differences in predictive performance



Fairness-Aware Machine Learning

Fairness as an optimization problem: technical approaches towards optimizing for fairness.

- **Preprocessing**. *Change the data* such that the sensitive feature cannot be deduced from the data, but all other information is preserved as best as possible.
- **Constrained learning ('in-processing')**. Incorporate *explicit fairness constraints* in the learning algorthm.
- **Post-processing**. *Adjust* an existing machine learning model or its predictions such that it adheres to fairness constraints.

Conclusion

- Machine learning systems can reproduce, amplify, and introduce **unfairness**.
- There are different types of **fairness-related harms**, today we will focus on:
  - **allocation harm**
  - **quality-of-service harm**.
- Fairness-related harms can arise due to a **mismatch** between what we **optimize** for and what we actually **value**.
- **Fairness metrics** can be used to measure potential fairness-related harms.
- **Fair-ml algorithms** can be used to optimize for fairness metrics.

```
if questions:
    try:
        answer()
    except RuntimeError:
        pass
else:
    print("Thank You.")
```

Tutorial
Pre-Trial Risk Assessment

Tutorial
Measuring Group Fairness in Pre-trial Risk Assessment

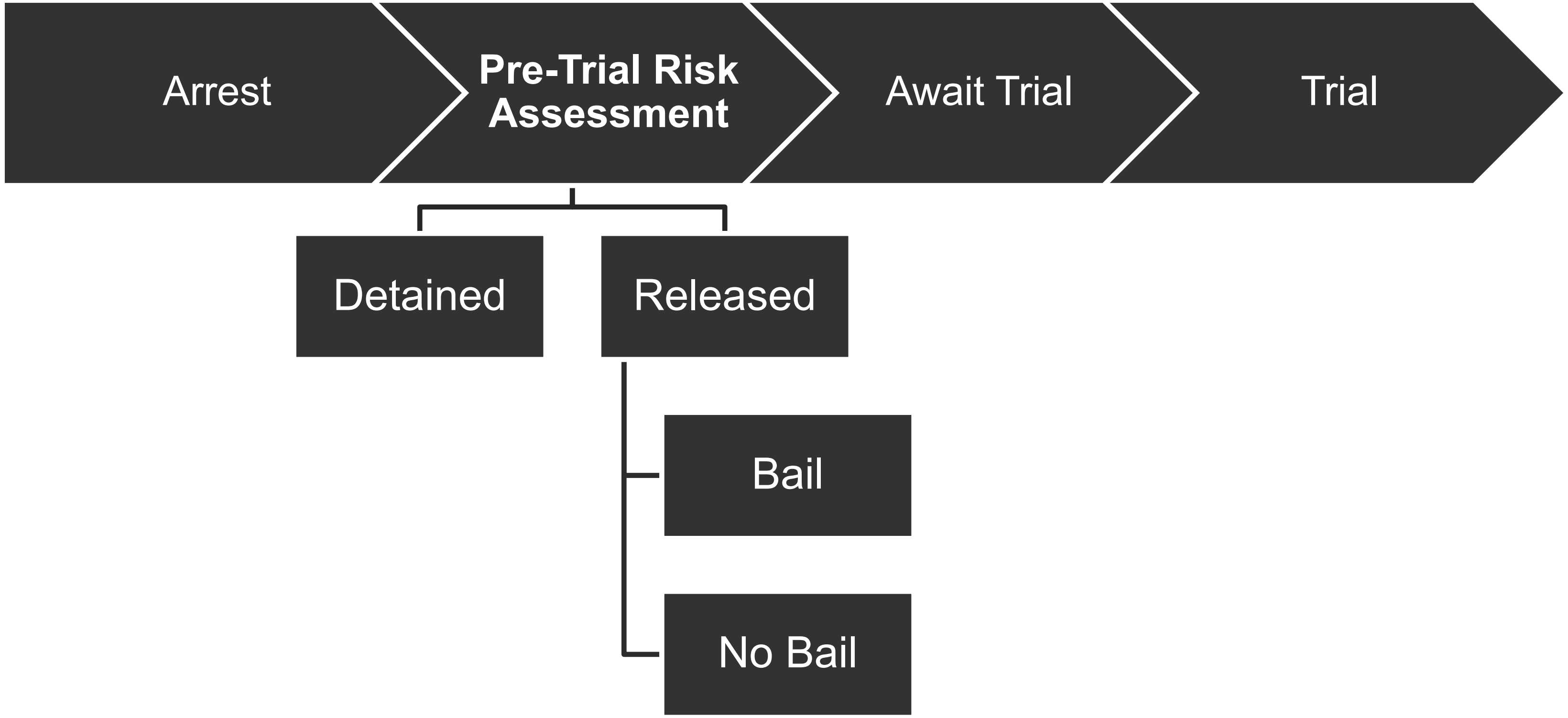
Tutorial.

- **Goal**: Measure notions of fairness using Fairlearn
- **Case study**: COMPAS recidivism prediction

Learning Objectives. After completing this tutorial, you will be able to:

- apply **group fairness metrics** in Python;
- explain several **trade-offs** between different group fairness criteria;
- explain how threats to **construct validity** may impact downstream **fairness-related harms**;

Context Pre-trial Risk Assessment in the United States



Propublica’s Analysis of COMPAS

In 2017, Propublica found that **COMPAS**, a recidivism prediction algorithm used by judges in the United States, failed differently for African-American defendants compared to white Americans.

	White	African-American
False Positive	23.5%	44.9%
False Negative	47.7%	28.0%

Source: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing