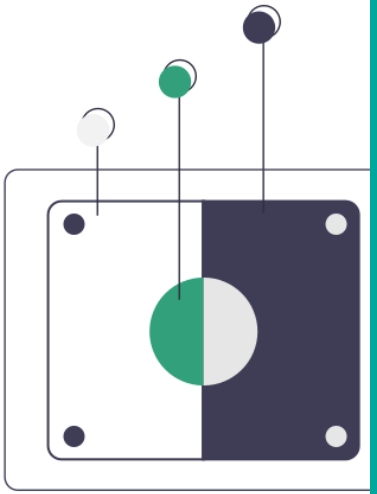
<AI & EQUALITY> A Human Rights Toolbox

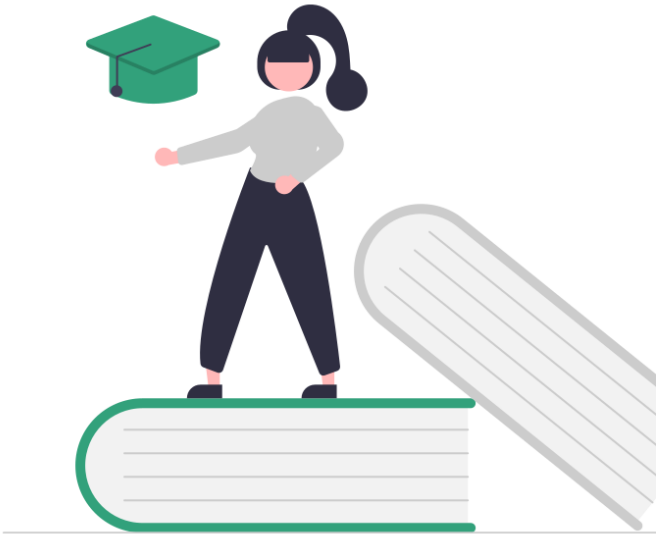
https://aiequalitytoolbox.com

Sofia Kypraiou (EPFL, Women at the Table)
Caitlin Craft-Buchman (Women at the Table, A+ Alliance)
Prof. Daniel Gatica-Perez (EPFL, Idiap)

3 September 2021



<AI & Equality>: A Human Rights-Based Approach

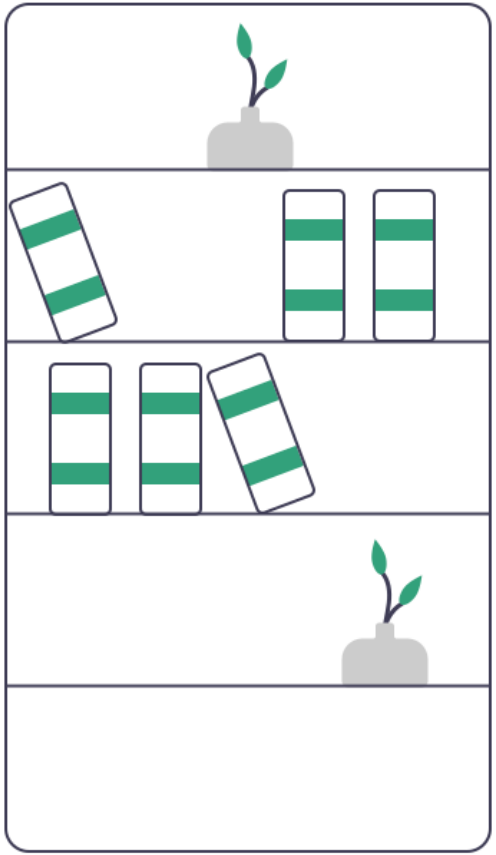


- Methodology includes:
  - Workshop,
  - outreach,
  - community plan

- Incorporate Human Rights concepts and data science

- A joint work between **EPFL**, **Women at the table** and in collaboration with the **Office of the United Nations High Commissioner for Human Rights**

- Goal: Integrate **International Human Rights frameworks** with current concepts of **fairness** for the design of an **educational tool** for **computer scientists**



MOTIVATION

Bias in algorithms

MOTIVATION

Bias in algorithms

Social impact of algorithms to our lives

MOTIVATION

Bias in algorithms

Social impact of algorithms to our lives

Companies use ethics as a way to address these concerns

Sofia Kypraiou

MOTIVATION

- Bias in algorithms
- Social impact of algorithms to our lives
- Companies use ethics as a way to address these concerns
- Data Science practicioners lack the connection between data science and society

MOTIVATION

- Bias in algorithms
- Social impact of algorithms to our lives
- Companies use ethics as a way to address these concerns
- Data Science practicioners lack the connection between data science and society
- Universities have this role to train computer scientists

Tech Ethics in universities

Interdisciplinary approaches
> 300 courses in different universities

- Viewed as a **specialization,** something that someone else does

- **stand-alone** and **unconnected** from computer science education so student's don't see the importance

- **barriers** come from **interdisciplinarity**.
  - Different vocabulary between technologists and philosophers

Sofia Kypraiou

What are Human Rights?

"Human rights are **rights** we have **simply because we exist as human beings** - they are not granted by any state.

These universal rights are **inherent** to us all, regardless of nationality, sex, national or ethnic origin, color, religion, language, or any other status.

They range from the most fundamental the **right to life** - to those that **make life worth living**, such as the rights to food, education, work, health, and liberty"

-- (The United Nations, *"Universal Declaration of Human Rights"*, 1948).

Why Human Rights?

- often better **defined** and **measurable**
- most are defined under **international or national law.**
- converts **voluntary promises** of ethical behaviour into **compulsory requirements** for compliance with established legislation.
- exceeds **national** and **cultural borders**


Article 1
Free and equal
Article 2
Freedom from discrimination
Article 3
Right to life
Article 4
Freedom from slavery
Article 5
Freedom from torture
Article 6
Right to recognition before the law
Article 7
Right to equality before the law
Article 8
Access to justice
Article 9
Freedom from arbitrary detention
Article 10
Right to a fair trial
Article 11
Presumption of innocence
Article 12
Right to privacy
Article 13
Freedom of movement
Article 14
Right to asylum
Article 15
Right to nationality
Article 16
Right to marriage and to found a family
Article 17
Right to own property
Article 18
Freedom of religion or belief
Article 19
Freedom of expression
Article 20
Freedom of assembly
Article 21
Right to partake in public affairs
Article 22
Right to social security
Article 23
Right to work
Article 24
Right to leisure and rest
Article 25
Right to adequate standard of living
Article 26
Right to education
Article 27
Right to take part in cultural, artistic and scientific life
Article 28
Right to a free and fair world
Article 29
Duty to your community
Article 30
Rights are inalienable
B-Tech

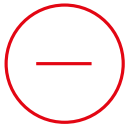
Sofia Kypraiou

Human Rights-Based Approach (HRBA)

- **Conceptual framework** without a universal definition
- Different actors use slightly different version of a HRBA depending on **context**

A HRBA should:

- comply with **the human rights law**,
- further the realisation of **human rights**



- Understood primarily on **legal terms**
- **Not translated** into general **guidance** for data scientists

Sofia Kypraiou

<AI & Equality> Methodology

"How to apply a Human Rights-Based Approach (HRBA) to AI"

Material:
https://aiequalitytoolbox.com/resources.html

- Data/computer science students
- 3h workshop
- Using Jupiter notebook to integrate theory with practice

<AI & Equality> Learning outcomes

- 1. Explain **a human rights-based approach** to **AI**
- 2. **Identify** the relevance of **different biases** and importance of intersectionality, gender equality and bias **to computer science** and engineering / institutional objectives
- 3. **Analyse** how gender, racial and other **bias has occurred or can occur** in the research, design and development of **AI**
- 4. **Apply** how and when to use tools and techniques **to mitigate bias** in AI
- 5. Evaluate methods to **integrate non-discrimination** into design, planning and implementation of **AI** projects

Workshop structure

Sofia Kypraiou

Part I

A. Human Rights Module

B. Applied Research

Part II

A. Practical Toolbox

Workshop structure

Part I

A. Human Rights Module

- Taught by **human rights / legal experts**
- Introducing basic human rights concepts which are relevant to the designing of algorithms
- and a Human Rights-Based Approach (**HRBA**) to machine Learning
- Focuses on human rights principles

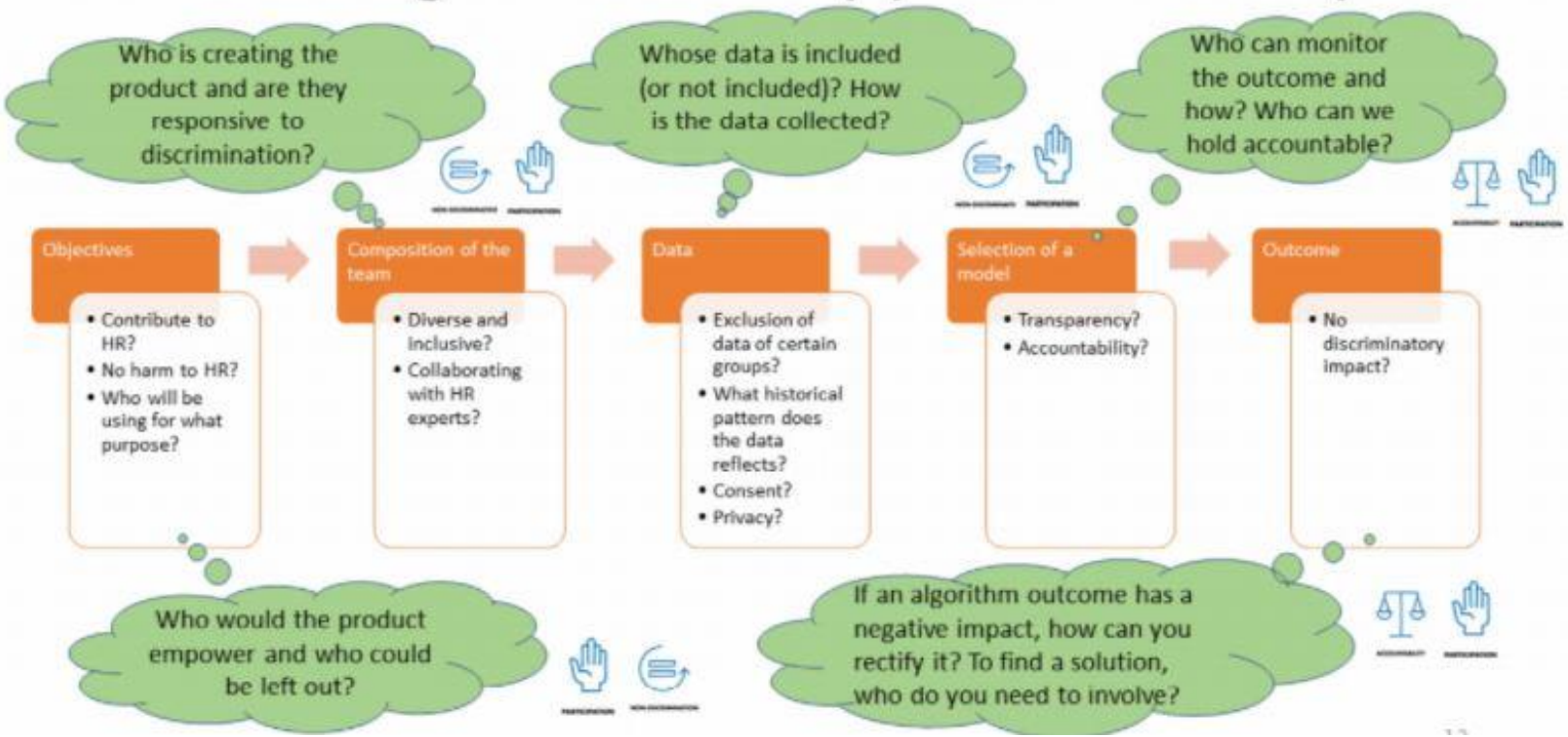
Part I A. Human Rights Principles

NON-DISCRIMINATION

PARTICIPATION

ACCOUNTABILITY

A human rights-based approach: example



Workshop structure

Part I

A. Human Rights Module

B. Applied Research

- Research Representatives (**PhD students, post-doc, faculty**) review the social impact of algorithms
- participants connect human rights and current research
- presenters showcase their work linked between AI and human rights

Sofia Kypraiou

Workshop structure

Part II

- **Step-by-step case study**

- apply Human Rights-Based Approach (HRBA) in practice (debiasing data and algorithms)

- **experiment** with data to see how different mathematical and data concepts of fairness interrelate
- begin a **critical analysis checklist** of the data process
- apply some of the concepts and debiasing literature to hands-on exercise

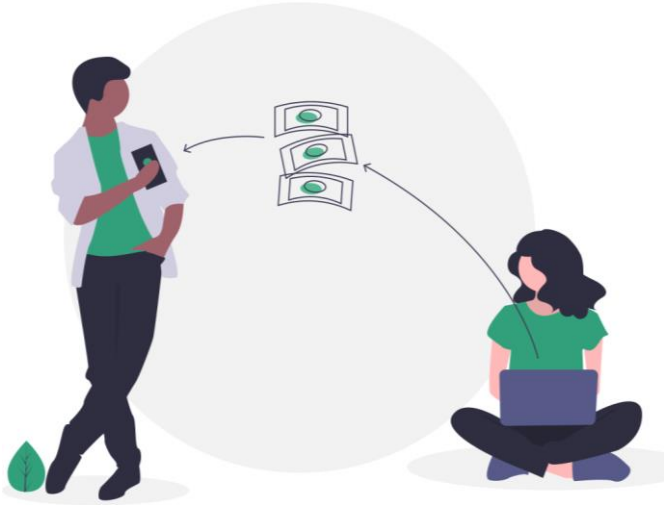
A. Practical Toolbox

Part II A. Practical Toolbox: Use case

- **loan application**
- predict if an applicant will be able to repay a loan according to the set of attributes available in the dataset

- Use (group and individual) **fairness metrics** to measure **gender equality**

- Explore different sources of **bias** and ways to **mitigate** it



Part II A. Practical Toolbox: Structure

INTRODUCE FAIRNESS METRICS

PERFORM EXPLORATORY DATA ANALYSIS AND CREATE A BASELINE MODEL

(PRE-PROCESSING) REBALANCE THE DATA

(IN-PROCESSING) BUILD A MODEL WITH FAIRNESS CONSTRAINTS USING A META CLASSIFIER

(POST-PROCESSING) OPTIMISE FOR THE DIFFERENT GROUPS OF PEOPLE USING EQUALISED ODDS

Part II A. Practical Toolbox: Fairness

INTRODUCE FAIRNESS METRICS

Fairness is:

- Is **not technical**, but **ethical** concept
- is **contextual**, no one-size-fits-all approach
- Has **no set answers,** often cost/benefit decisions have to be made
- **Process**, no single fairness checkpoint

Part II A. Practical Toolbox: Baseline model

Sofia Kypraiou

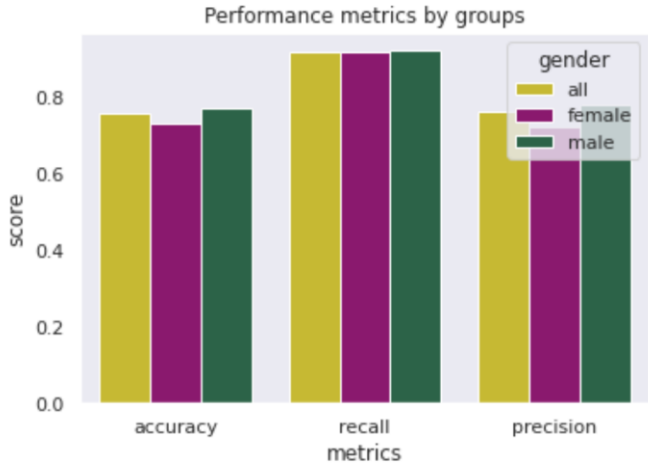
PERFORM EXPLORATORY DATA ANALYSIS AND CREATE A BASELINE MODEL

Questions ("Datasheets for Datasets", Gebru et al., 2018)

Why was the dataset created?
Does the dataset identify any subpopulations (e.g., by age, gender)? how these subpopulations are identified?

Exploratory Data Analysis

- Importance of evaluating model for different subgroups

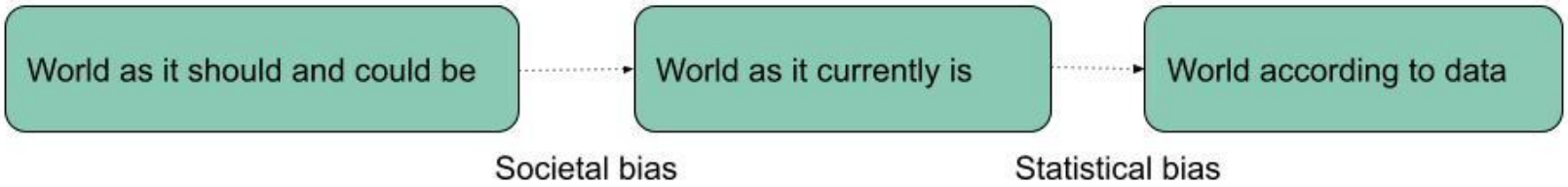


SpeSofia Kypraiou
aker

Part II A. Practical Toolbox: Pre- processing (data)

(PRE-PROCESSING)

- 1. **Where** bias exists in data & Different **types** of data biases
  - Statistical bias
  - Societal bias
- 2. Methods to **mitigate** bias

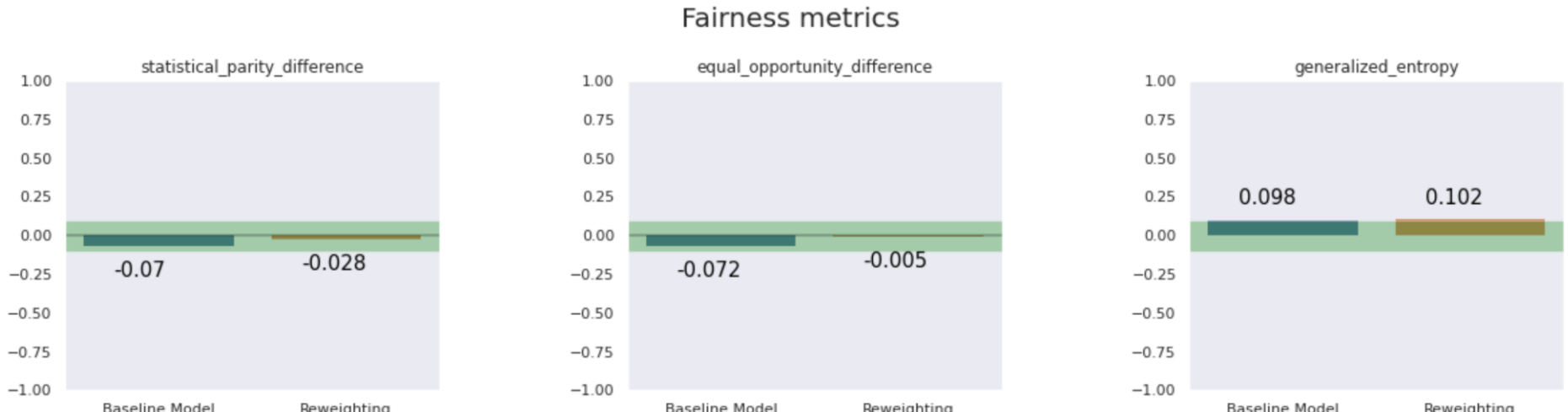


SpeSofia Kypraiou
aker

Part II A. Practical Toolbox: Pre- processing (data)

(PRE-PROCESSING)

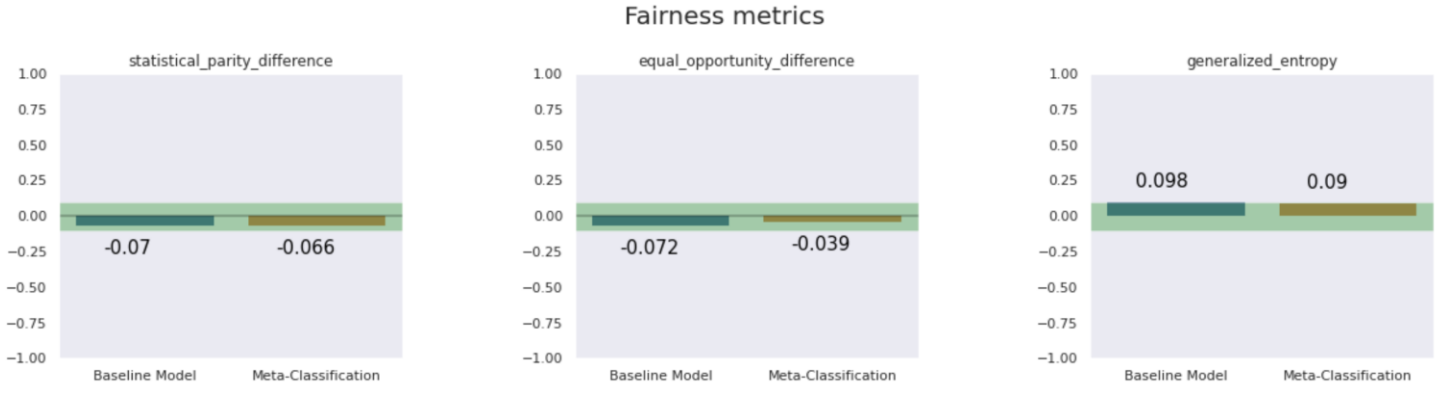
- pre-processing technique: Reweighting algorithm (Kamiran & Calders, 2011)



Part II A. Practical Toolbox: In- processing (model)

(IN-PROCESSING)

- 1. **bias** from:
  - inappropriate data handling
  - inappropriate model selection
  - incorrect algorithmic design or application
- 2. Methods to **mitigate** bias
  - Explainability, Covariate selection

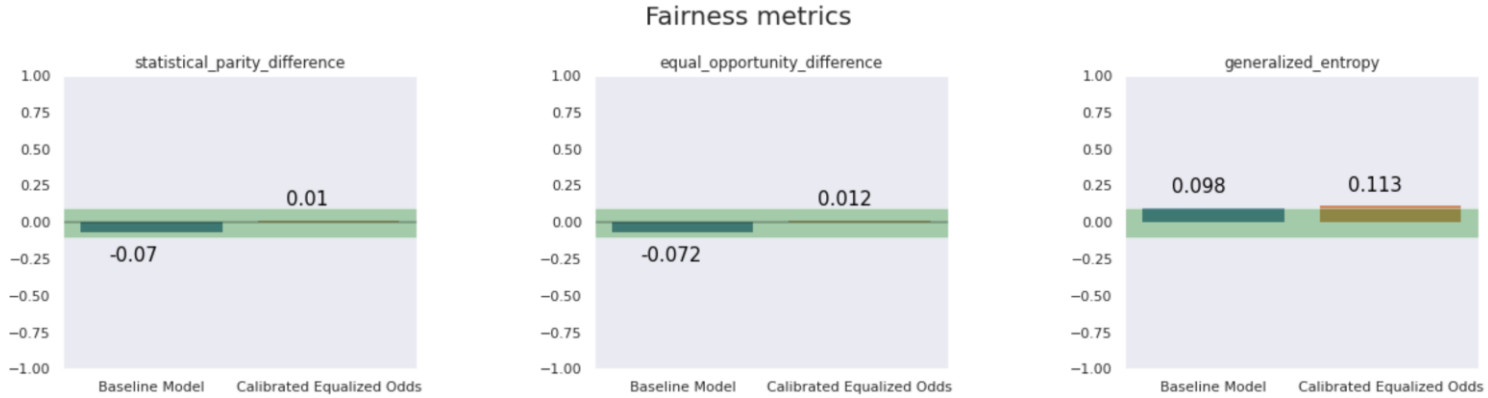


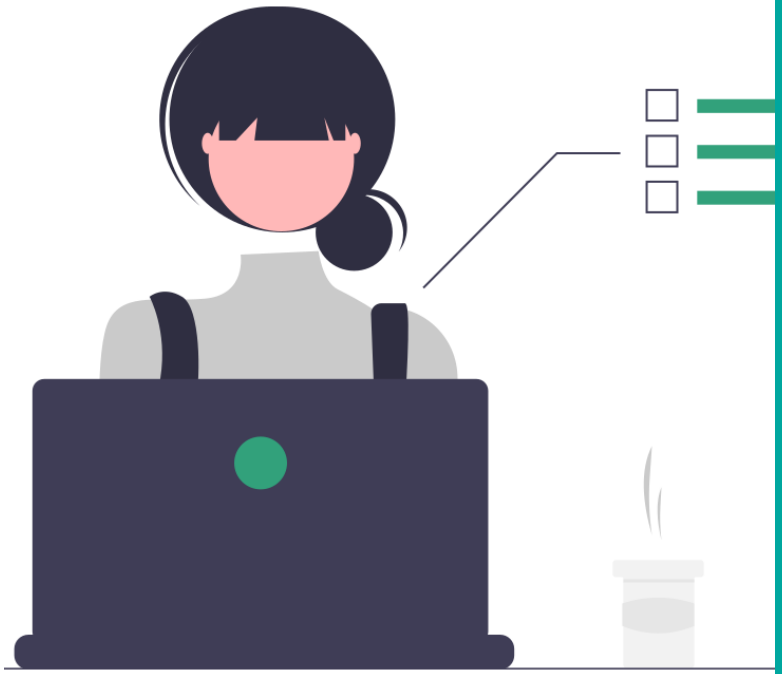
SpeSofia Kypraiou
aker

Part II A. Practical Toolbox: Post- processing (predictions)

(POST-PROCESSING)

- Assumptions when evaluating decisions:
  - Decisions are **evaluated as an aggregation** of separately evaluated individual decisions
  - All individuals are considered **symmetrically**
  - Decisions are evaluated **simultaneously**





Sofia Kypraiou

École polytechnique fédérale de Lausanne (EPFL) Workshop

- Evaluation procedure:
  - Questionnaire
  - Semi-structured interview with participant, Feedback from colleagues

University College Dublin (UCD) Workshop

- 20 May 2021
- 12 final participants

- Differences:
  - Reduced to **2-hours**
  - **Removed applied** research
  - EPFL: data science participants
  - UCD: social science participants

Objectives

	Objective	Before workshop (/5)	After workshop (/5)	Increase (%)
1.	I would rate my confidence in describing the key elements of a human rights based approach to AI	2.5	4.5	80
2.	I would rate my confidence in describing the equality and gender, racial and other forms of discrimination relevant to the design of algorithms	3.125	4.5	44
3.	Overall, I would rate my ability to identify the relevance of different biases and the importance of gender, race and equality to computer science and engineering	3.75	4.75	26
4.	I would rate my ability to analyse how gender, racial and other bias has occurred or can occur in the research, design and development of AI	3.25	4.25	30
5.	I would rate my ability to use tools and techniques to mitigate bias in AI	2.375	3.625	52
6.	I would rate my ability to evaluate methods to integrate non-discrimination into design, planning and implementation of AI projects	2.5	4	60

Overall satisfaction: 4.7/5 = 98%

Summary

Contribution

Limitations

Future work

Contribution

Collaboration with Office of the United Nations High Commissioner for Human Rights (OHCR)

Transforming a human-rights based approach into actionable steps for data scientists

Limitations

- Evaluation and validation: hard to capture to what degree students learn the critical concepts

Limitations

Sofia Kypraiou

Evaluation and validation

Content and Material: rich yet quite heavy

Sofia Kypraiou

Limitations

- Evaluation and validation
- Content and Material
- Organisation of workshops: Human resources and organisation

Future work

Blended workshops

- call the academic community to **add more legal/ethical/social science material** and help understand **how code can** potentially create **insights and solutions**

- academics from a **technical** and **law/human rights/social sciences background** to perform the blended workshops **jointly**.

Future work

Sofia Kypraiou

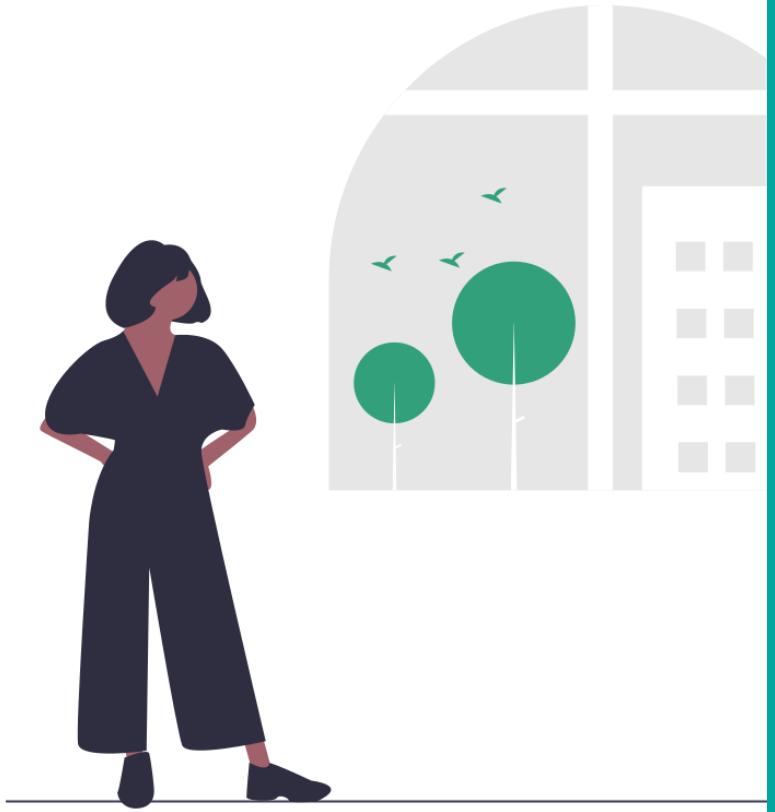
Blended workshops

Community outreach

- **international community** of university that want to make a difference.
- technical and social scientists to help foster **multi-disciplinary collaboration** on this cross cutting topic.

Future work

- Blended workshops
- Community outreach
- Validation of the methodology



01. Introduction
02. Background
03. Methodology
04. Workshops
05. Discussion
06. Conclusion

“ To **respect** these **(human) rights** in our rapidly evolving world, we must ensure **that the digital revolution is serving the people**, and not the other way round. We must ensure that **every machine-driven process** or artificial intelligence system complies with cornerstone principles such as **transparency**, **fairness**, **accountability**, oversight and redress.”

- -- Michelle Bachelet, the UN High Commissioner for Human Rights, keynote speech for “Human rights in the digital age” (Bachelet, 2019)

Let’s collaborate!

Contact:
sofia@womenatthetable.net

THANK YOU!
https://aiequalitytoolbox.com/