# FAIRNESS

**@Responsible ML Winter School**

Lili Jiang, associate professor
Department of Computing Science


UMEÅ UNIVERSITY

# ME

- PhD in computer science
- Research experience/interest
  - Data science (e.g., information extraction, entity-based knowledge graph, information retrieval et c),
  - AI trustworthiness (e.g., fairness, privacy).
- Homepage: https://people.cs.umu.se/ljiang/

# TODAY

- AI Fairness
  - AI Fairness in a nutshell
  - Definitions and metrics
  - Bias in AI systems' life cycle
  - AEQUITAS project (a glance)

UMEÅ UNIVERSITY

# TODAY

- AI Fairness
  - AI Fairness in a nutshell
  - Definitions and metrics
  - Bias in AI systems' life cycle
  - AEQUITAS project (a glance)

# TRUSTWORTHY AI [1]

AI system should be:

- Lawful
- Ethical
- Robust

# REQUIREMENTS OF TRUSTWORTHY AI

Ensures that the development, deployment and use of AI systems meets the seven key requirements for Trustworthy AI:

(1) human agency and oversight,

(2) technical robustness and safety,

(3) privacy and data governance,

(4) transparency,

(5) diversity, non-discrimination and fairness,

(6) environmental and societal well-being,

(7) accountability.

# WHAT IS FAIRNESS

- Fairness: Impartial and just treatment or behaviour without favouritism or discrimination.
- AI fairness: AI systems should treat all people fairly.

- The other side of the coin:

  Unfairness

  Bias

  Discrimination

# SOME INFAMOUS EXAMPLES

- COMPAS system used by US courts predicts higher values to the black defendents than their actual risks.
- AI-driven diagnostic tools for skin cancer are less accurate for individuals with dark skin.
- Amazon's automated recruiting tools was found to be biased agains women.
- Digital Ageism: tech companies design apps and websites with small fonts, complex navigation.
- .....

# GDPR (FAIRNESS AND EXPLAINABILITY)

Article 5(1) requires that personal data shall be: “(a) processed lawfully, fairly and in a transparent manner in relation to individuals ('lawfulness, fairness and transparency’);

Articles 21 and 22 suggest that the right to understand “meaningful information” about and the “significance” of, automated processing is related to an individual’s ability to opt out of such processing.

# CHALLENGES/TRADEOFFS

- Bias is as old as human civilization.
- Intentional and indirect discrimination both exist.
- Bias exists in data/model/validation etc. in the whole AI life-cycle.
- Different contexts/applications/stakeholders
- General lacking of regulations.
- No consensus on definition, state-of-the-art methodologies, metrics etc.
- Requires interdisciplinary efforts.
- ……

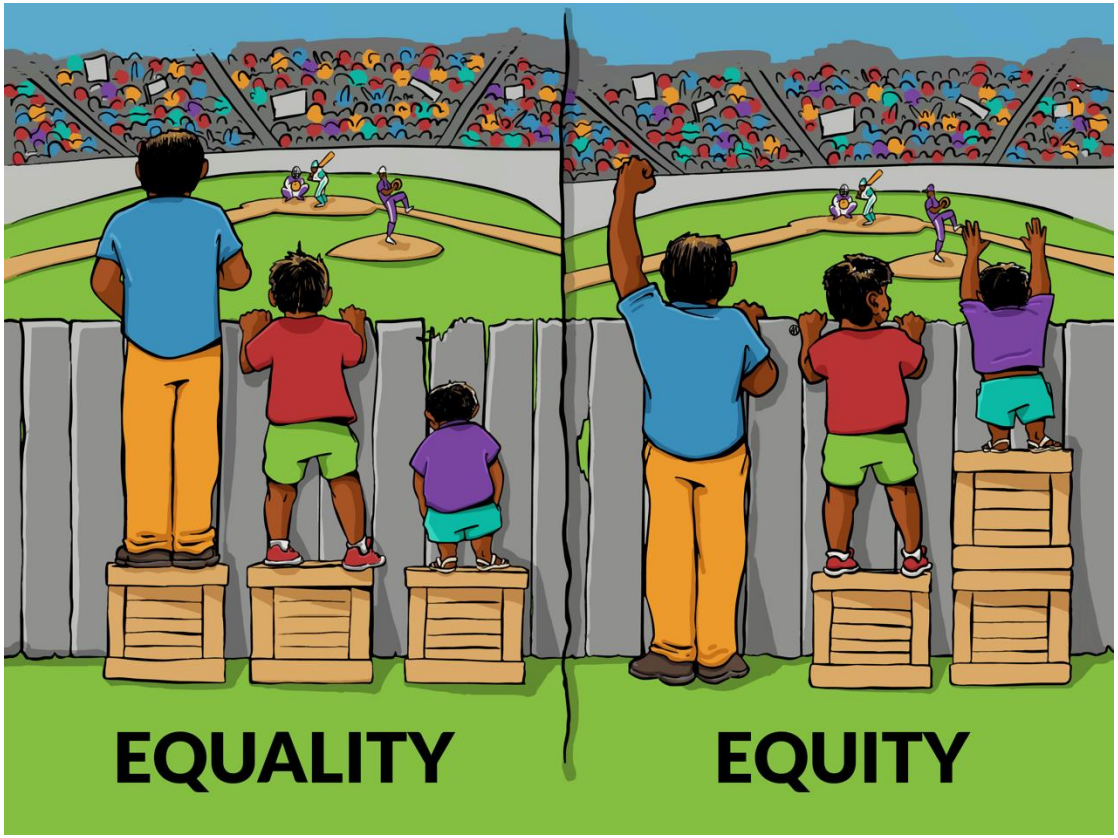
# TODAY

- AI Fairness
  - AI Fairness in a nutshell
  - Definitions and metrics
  - Bias in AI systems' life cycle
  - AEQUITAS project (a glance)

# **FAIRNESS** - DEFINITIONS


NO one defintion of fairness applicable in all contexts

# EQUALITY VS EQUITY?




Interaction Institute for Social Change | Artist: Angus Maguire.

# EQUALITY VS EQUITY

- Equality = Everyone gets the same resources, opportunities, or treatment, regardless of their individual circumstances.

Example: AI screening tools in the recruitment procedure apply identical filters to all resumes, without adjusting for systemic disadvantages (e.g., education gaps or biased historical hiring trends).

- Equity = Everyone gets what they need to succeed, which may mean different levels of support for different people.

Example: Hiring teams apply contextual evaluation, recognizing that candidates from underprivileged backgrounds may have had fewer opportunities but still show strong potential.

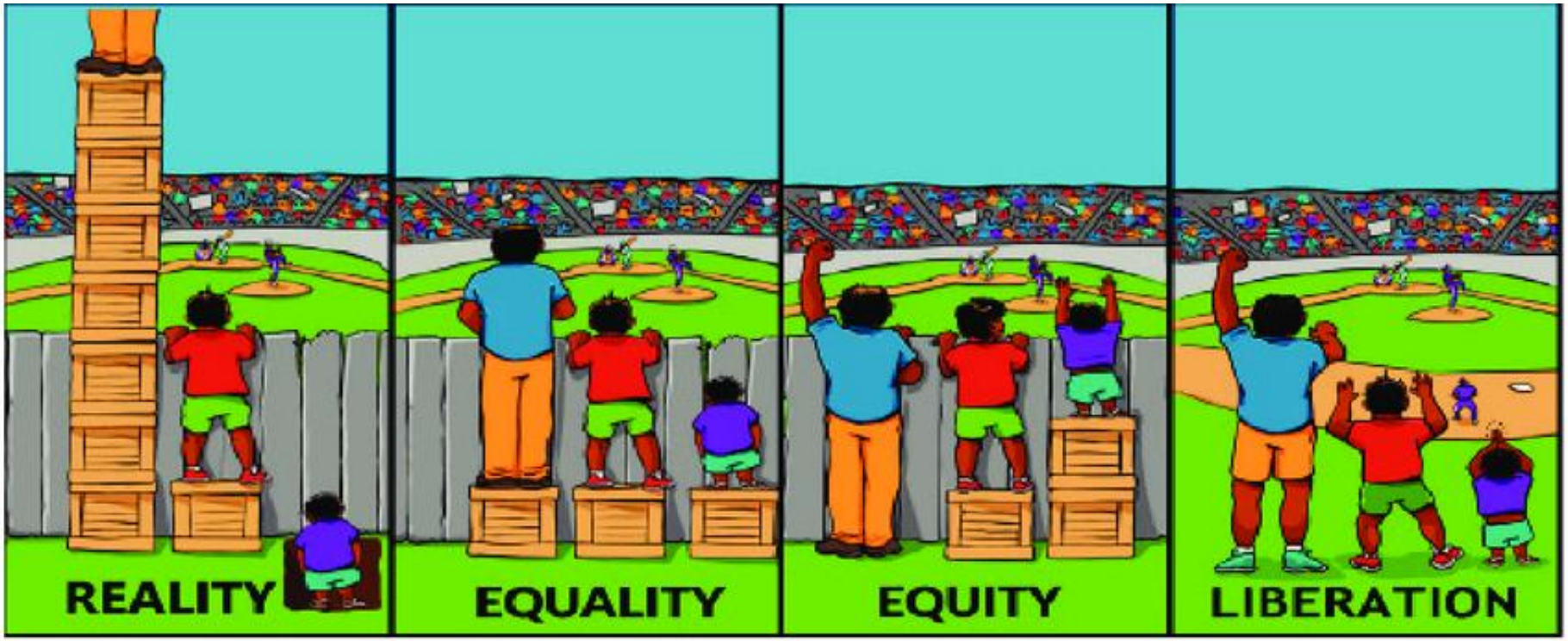
# AI FAIRNESS – NOTION [2]

Fairness: AI systems should treat all people fairly.

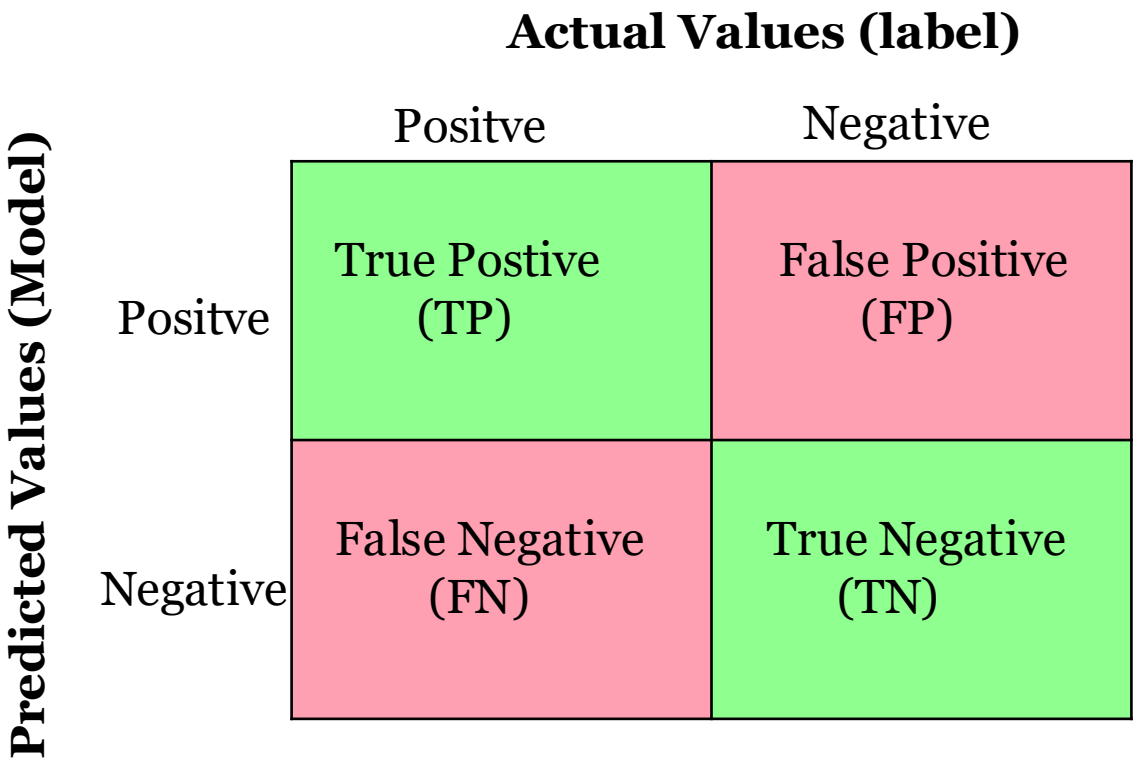
Fairness is most often conceptualized as equality of opportunity.

| | |
|---|---|
| Narrow view: | Ensure that people who are similarly qualified for an opportunity have similar chances of obtaining it. |
| Broad view: | Ensure people of equal ability and ambition are able to realize their potential equally well. |
| Middle view: | Discount differences due to past injustice that accounts for current differences in qualifications. |

# A FOLLOWUP



# CONFUSION MATRIX



# **FAIRNESS** – METRICS [5]

- Equalized odds and equality of opportunity
  - A predictor satisfies equalized odds if both the true positive rate (TPR) and (separately) the false positive rate (FPR) are the same across groups.
- Group fairness metrics
  - Disparate impact
  - Statistical parity difference
  - Equal opportunity difference
  - Demographic parity
- Predictive parity: is satisfied when the positive predictive value(PPV) is the same for both groups.
- Calibration: an algorithm is calibrated if for all scores, the individuals who have the same score have the same probability of belonging to the positive class, regardless of group membership.

See reference [5] for more on these metrics.

# EQUALIZED ODDS
## (HIRING EXAMPLE)

- The goal of the **equalized odds fairness metric** is to ensure a machine learning model performs equally well for different groups.

TPR = TP/ (TP+FN)

FPR = FP/ (FP + TN)

In the hiring example:

TPR = the probability for a qualified applicant being hired. (correctly beneficial)

FPR = the probability of an unqualified applicant received offer. (wrongly beneficial)

To be fair ML system, they should be equal across groups:

TPR(protected) = TPR(unprotected)

FPR(protected) = FPR(unprotected)

# TODAY

- AI Fairness
  - AI Fairness in a nutshell
  - Definitions and metrics
  - Bias in AI systems' life cycle
    - Bias discovery
    - Bias Mitigation
    - Causal fairness
  - AEQUITAS project (a glance)

UMEÅ UNIVERSITY

# **FAIRNESS** - BIAS DISCOVERY [3]

- Bias in Data
  - Historical bias
  - Representation bias
  - Measurement bias
  - .....
- Bias in Modeling
  - Aggregation bias
  - Evaluation bias
  - .....

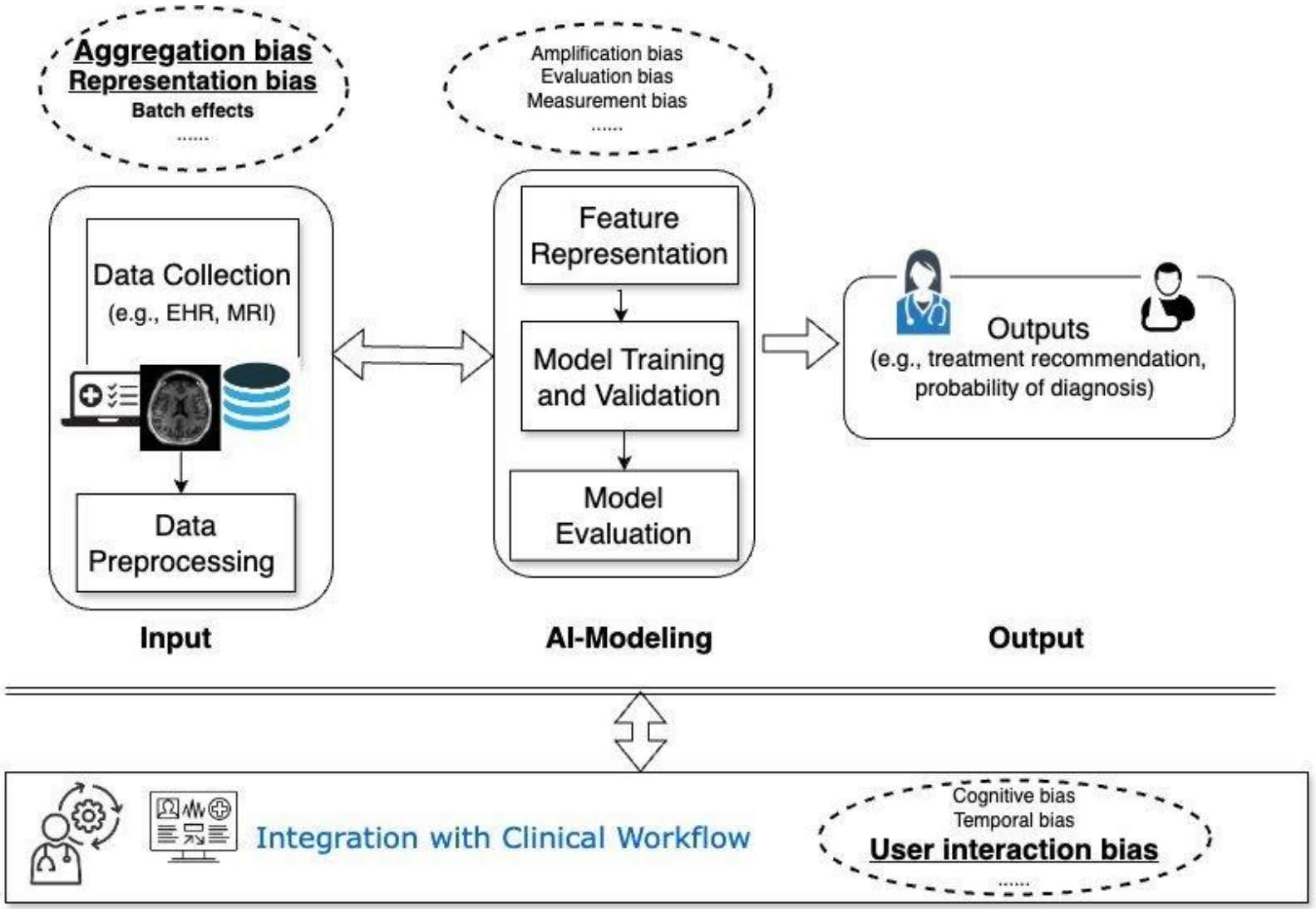
# BIAS IN DATA

- Historical bias
  - E.g., "man is to computer programmer as woman is to homemaker,"
- Representation bias
  - Underrepresentation bias (e.g., less data from lower-income background).
  - Overrepresentation bias (e.g., more white faces in image dataset)
- Measurement bias
  - Data that's easily available is often a noisy proxy for the actual features or labels of interest
  - E.g., black defendants getting harsher sentences than white defendants for the same crime.

# BIAS IN MODELING

- Aggregation bias
  - Distinct populations are inappropriately combined, but a single model is unlikely to suit all groups.
  - E.g., in health care, some models used the disesese levels differ in complicated ways across ethnicities.
- Evaluation bias
  - A model is optimized using training data, but its quality is often measured against certain benchmarks, which do not represent the general population.
  - E.g., the evaluator assesses individuals based on data about their past education and employment.

# POSSIBLE BIAS IN AI SYSTEM (E.G., HEALTHCARE )



# AGGREGATION BIAS

This bias occurs when an inference is made about an individual based on their membership within a group. These unintentional weightings of certain factors can cause algorithmic results that exacerbate and reinforce societal inequities.

**Example**: Simpson's Paradox is a statistical phenomenon where a trend appears in different groups of data but disappears or reverses when the groups are combined. Suppose a hospital is evaluating the performance of an AI-based system for disease diagnosing. The hospital collects data from two different departments: A (primarily deals with early-stage cases) and B (handles more advanced cases).

In **Department A**, the AI accurately diagnoses **80%** of cases among male patients, **90%** of cases among female patients. In **Department B**, the AI accurately diagnoses **15%** of cases among male patients, **33%** of cases among female patients. However, when the data from **both departments** are combined and analyzed as a whole, they have **68%** of cases male patients and **56%** of cases among female patients.

**Life-cycle**: model development, validation and evaluation.

# USER INTERACTION BIAS

This arises when a user imposes their own self-selection biases and behavior during interaction with data, output, result. This Bias can be influenced by other subtypes of bias.

**Example**: Patients are asked to provide feedback on their experience and satisfaction with an AI-based skin condition diagnosis process. In this scenario, user interaction bias can occur if patients' feedback is influenced by their expectations or prior beliefs about the accuracy and reliability of AI technology in healthcare. The possible user interaction bias may include **positive confirmation bias** (patients who receive a diagnosis that aligns with their expectations or previous beliefs about their condition), **negative confirmation bias** (patients who have negative preconceptions or skepticism about AI technology in healthcare may be predisposed to providing negative feedback), and **feedback influence** (patients may be influenced by the feedback of other patients or external sources (e.g., online reviews, friends).

**Life-cycle**: whole AI decision system cycle

# FAIRNESS – BIAS MITIGATION [4]

- Pre-processing (on data)
- In-processing (on ML algorithm)
- Post-processing (on ML model)

See [3] for more details on these approaches

# PRE-PROCESSING METHODS

Produce a “balanced” dataset.

- Modifying the original data distribution by altering class labels of carefully selected samples close to the decision boundary.
- Assigning different weights to samples based on their group membership.
- Carefully sampling from each proteceted group.
- ….

How to know the protected groups??

# IN-PROCESSING METHODS

Reformulates the classification problem by explicitly incorporating the model's discrimination behavior in the objective function via regularization or constraints, or by training on latent target labels.

- Modifying the splitting criterion of decision trees to consider the impact of the protected attributes.
- Integrating a regularizer to reduce the effect of “indirect prejudice”.
- Redefining the classification problem by minimizing an arbitrary loss function subject to the individual fairness-constraint.
- Incorporating disparate mistreatment into logistic-regression and SVMs.
- ……

# IN-PROCESSING METHODS

For unsupervised learning:

- Fair-PCA approach forces equal reconstruction errors for both protected and unprotected groups.
- Fair clustering as having approximately equal representation for each protected group in every cluster and define fair-variants of classical k-means and k-medoids algorithms.

# POST-PROCESSING METHODS

Postprocesses the classification model once it has been learned from data.

- Differentiating the decision boundary itself over groups to keep proportionality of decisions among protected versus unprotected groups.
- Wrapping a fair classifier on top of a black-box base classifier.
- .....

# **FAIRNESS** – APPLICATION AND OPEN SOURCE

- **Fairlearn**: Fairlearn is a Python package that empowers developers of artificial intelligence (AI) systems to assess their system's fairness and mitigate any observed unfairness issues.
- **AI Fairness 360**: an. open-source toolkit of metrics to check for unwanted bias in datasets and machine learning models, and state-of-the-art algorithms to mitigate such bias.
- **FairML**: an end-to-end toolbox for auditing predictive models by quantifying the relative significance of the model's inputs.
- **Themis-ML**: an open-source machine learning library that implements several fairness-aware methods that comply with the sklearn API.
- ……

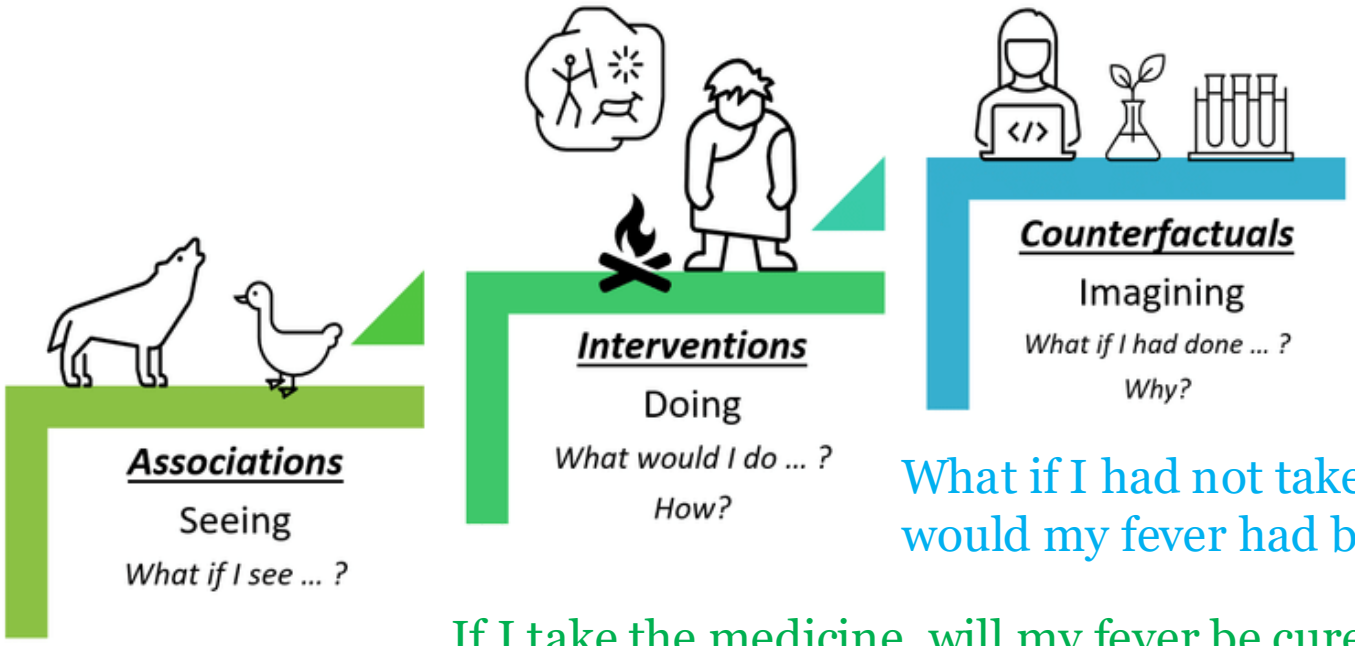
# TODAY

- AI Fairness
  - AI Fairness in a nutshell
  - Definitions and metrics
  - Bias in AI systems' life cycle
    - Bias discovery
    - Bias Mitigation
    - Causal fairness
  - AEQUITAS project (a glance)

UMEÅ UNIVERSITY

# CAUSAL FAIRNESS

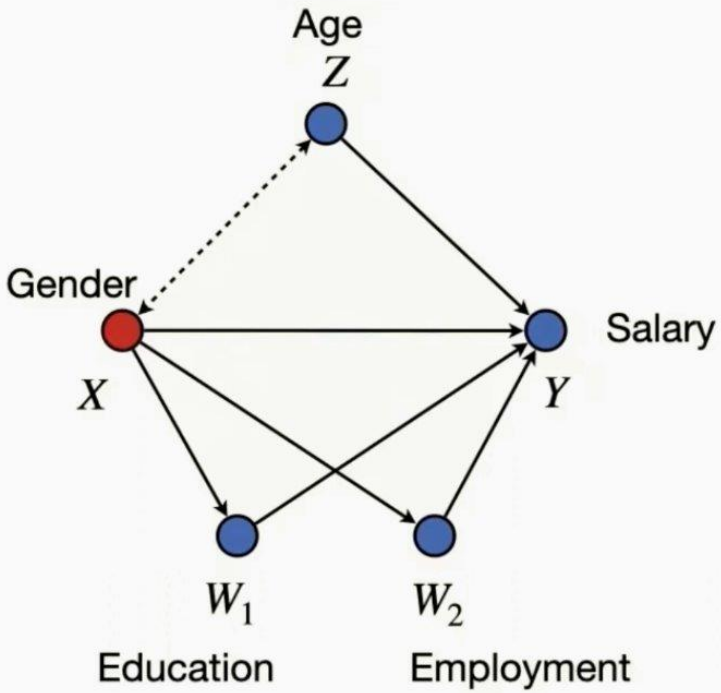
Pearl's Ladder of Causation




What if I had not taken the medicine, would my fever had been cured?

If I take the medicine, will my fever be cured?

What a symptom tell me about a disease?




DoWhy framework:

The **total effect** of Gender to Salary

We want to know exactly how the effect is created, so we want to split it up.

Causal Fairness Analysis:

- direct effect (DE):
  $X \to Y$
- indirect effect (IE):
  $X \to W_1 \to Y$ and $X \to W_2 \to Y$
- spurious effect (SE):
  $X \leftarrow Z \to Y$

Connection to the DoWhy framework:

- total effect: DE $-$ IE $-$ SE

# AN EXAMPLE

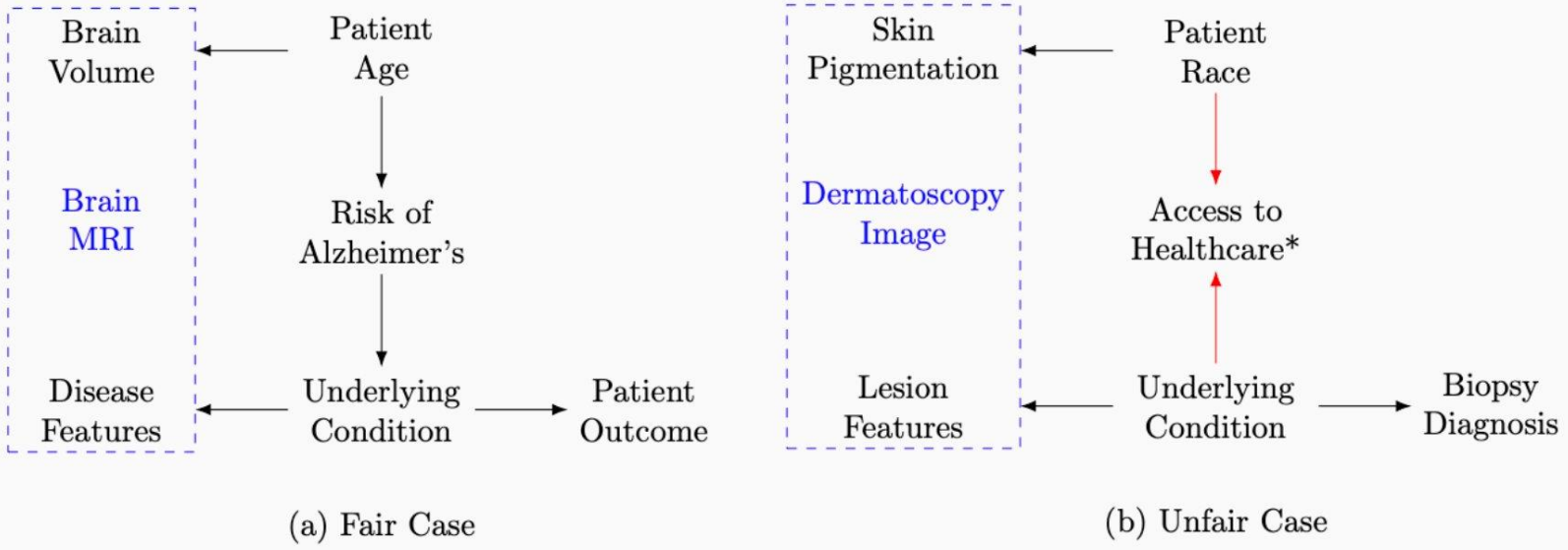
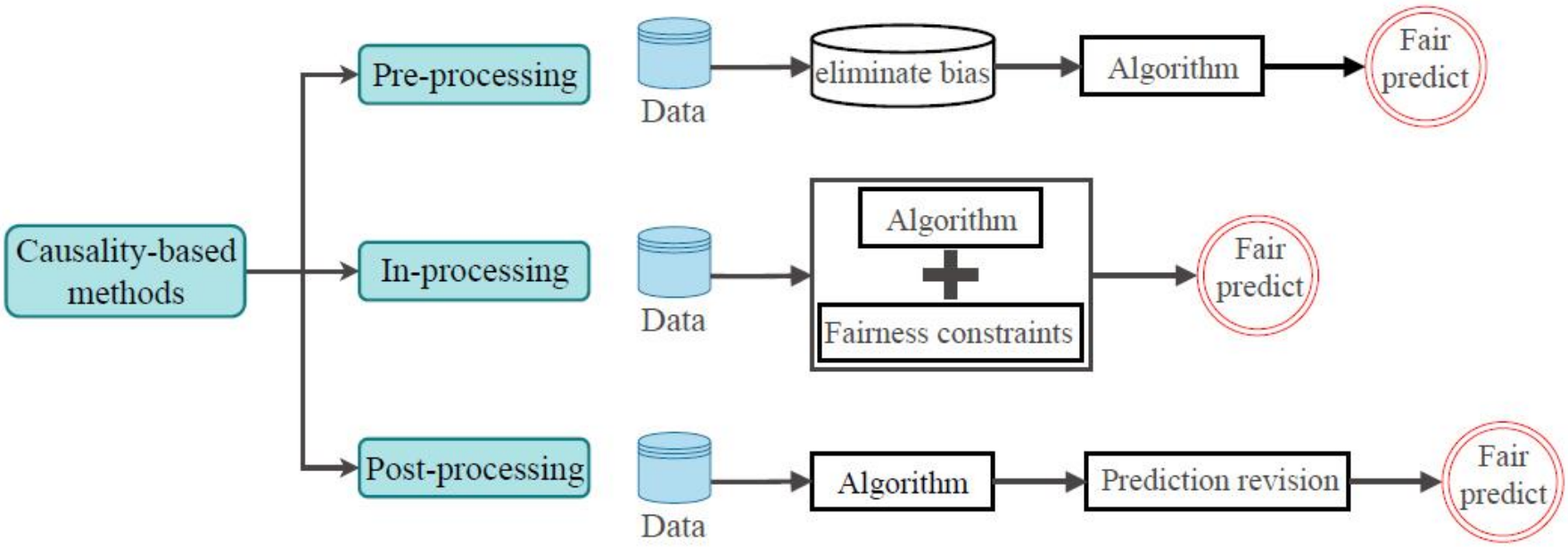



**Figure 1:** Example Causal Inference for medical Images (Jones et. al., 2023)

# CAUSAL FAIRNESS IN ML




From: A review of causality-based fairness machine learning

# REFERENCES

[1] European Commission, Directorate-General for Communications Networks, Content and Technology, *Ethics guidelines for trustworthy AI*, Publications Office, 2019, https://data.europa.eu/doi/10.2759/346720

[2] Solon Barocas and Moritz Hardt and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities,* fairmlbook.org, 2019.

[3] Suresh, H., & Guttag, J. (2021). *Understanding Potential Sources of Harm throughout the Machine Learning Life Cycle*. MIT Case Studies in Social and Ethical Responsibilities of Computing, (Summer 2021). https://doi.org/10.21428/2c646de5.c16a07bb

[4] Ntoutsi E, Fafalios P, Gadiraju U, et al. (2020). *Bias in data-driven artificial intelligence systems—An introductory survey*. WIREs Data Mining Knowl Discov. 2020;10:e1356.

[5] P. Garg, J. Villasenor and V. Foggo, "Fairness Metrics: A Comparative Analysis," *2020 IEEE International Conference on Big Data (Big Data)*, Atlanta, GA, USA, 2020, pp. 3662-3666, doi: 10.1109/BigData50022.2020.9378025.

# TODAY

- AI Fairness
  - AI Fairness in a nutshell
  - Definitions and metrics
  - Bias in AI systems' life cycle
  - AEQUITAS project (a glance)

UMEÅ UNIVERSITY

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement N° 101070363

AEQUITAS

# ASSESSMENT AND ENGINEERING OF EQUITABLE, UNBIASED, IMPARTIAL AND TRUSTWORTHY AI SYSTEMS

Call HORIZON-CL4-2021
Duration 1 November 2022 - 31 October 2025
Project id 101070363

UMEÅ UNIVERSITY

# CONSORTIUM


Universidad
de La Laguna


ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA


LOBA®


ALLAI.


P
PERIOD
think tank


UMEÅ UNIVERSITY


PHILIPS


ITI
INVESTIGATE
TO INNOVATE


EUROCADRES


UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh


Adecco


modis


ASOCIACION
RAYUELA


ARCIGAY
Associazione LGBT italiana


SERVIZIO SANITARIO REGIONALE
EMILIA-ROMAGNA
Azienda Ospedaliero - Universitaria di Bologna
IRCCS Istituto di Ricovero e Cura a Carattere Scientifico


POLICLINICO DI
SANT'ORSOLA

# AMBITION & OBJECTIVES

Address and tackle the multiple manifestations of bias and unfairness in AI proposing an

***open controlled experimentation environment***

for AI stakeholders to ***test fairness*** dimensions via controlled experiments

# AMBITION, OBJECTIVES & IMPACT

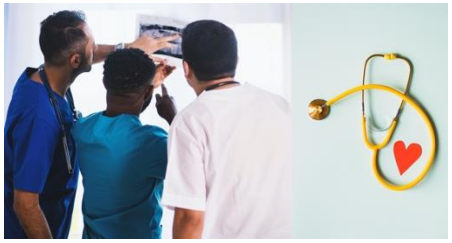
**Increase availability and deployment of unbiased AI solutions** releasing AEQUITAS both as a service via the AI-on-demand platform and as an on-premise tool

**Enhancing digital equality** and **social inclusion** for groups at risk of discrimination via new AI solutions

**Increase awareness**, knowledge, and skills about trustworthy, bias-free and **socially responsible AI** in the industry and scientific community

Reduce the gap between EGTAI and their adoption and implementation in private and public settings, grounding them into **practical methodologies**, **measurable KPIs** and **technical requirements**

# AMBITION & OBJECTIVES: USE CASES



## healthcare

- Fair tool supporting the diagnosis phase in **pediatric dermatology**
- **Bias-aware versions of ICU predictive algorithms** with higher chances of generalizing well


## human resources

- **Fair AI assisted recruiting system** to target the cognitive and structural bias associated with the recruiting process
- Assess possible bias in the human expert in selection of candidates

## social disadvantaged group

- Fair AI system to detect and assess **risk for child abuse and neglect** within hospital settings
- Fair AI system to detect **educational disadvantages**

# WWW.AEQUITAS-PROJECT.EU

# TAKE-AWAY ON AI FAIRNESS

- Bias in, bias out.
- Bias is along with human civilization and the whole life cycle of AI system.
- Bias cannot permanently removed.
- Fairness is not only about equality of opportunities.
- It is an interdiscipliary topic (norm, law, technology.)
- @YOU?

THANKS!

UMEÅ UNIVERSITY