Biased?

Dr. Marcel Blattner
Chief Data Scientist @Tamedia

Director of Studies
CAS Disruptive Technologies @HWZ

Terminology

- Artificial Intelligence

Schools of thoughts:

Thinking humanly
Cognitive approach. Requires to determine how humans think - cognitive revolution.

Thinking rationally
Laws of thoughts. Codify thinking with logic.

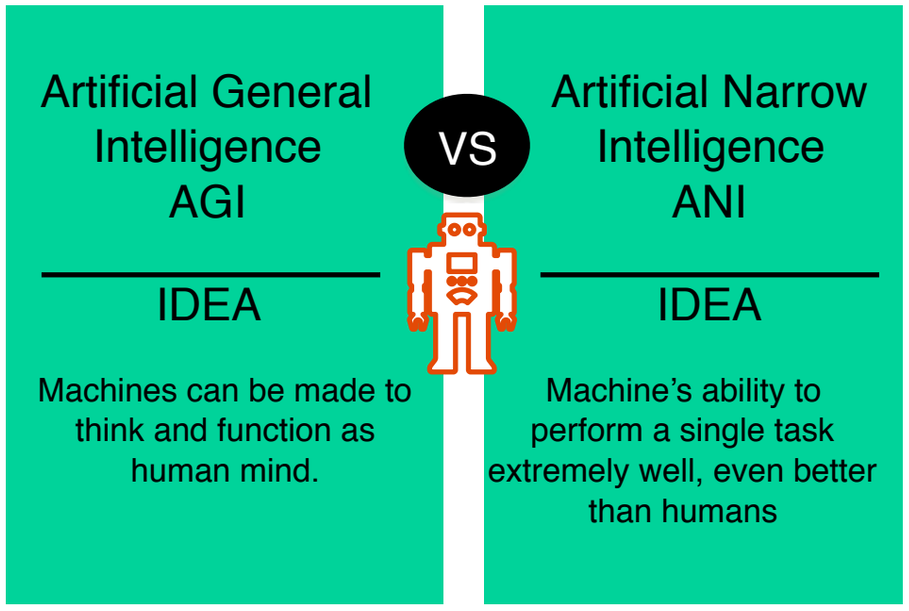
Acting humanly
Create a machine which acts like a human. This is called the Turing-Test approach. Needs computer vision and robotics.

Acting rationally
Maximise a goal achievement, given the available information. A rational agent is one that acts so as to achieve the best outcome.

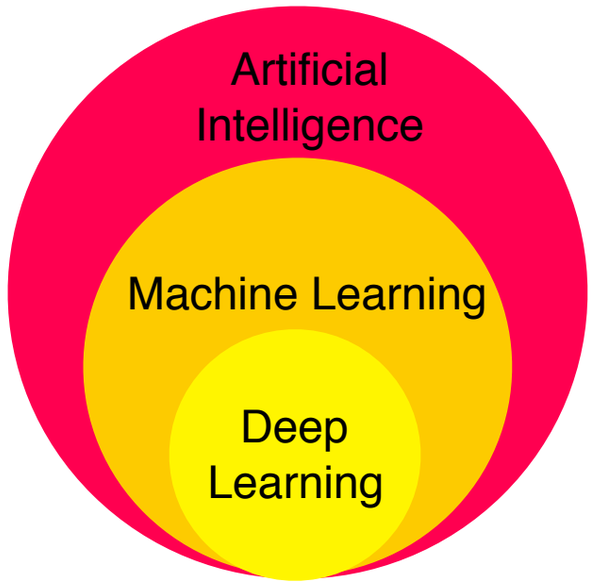
AI: The designing and building of intelligent agents that receive percepts from the environment and take actions that affect that environment.
Source: Russell&Norvig

AI is an interdisciplinary field of research!!

AGI vs. ANI:

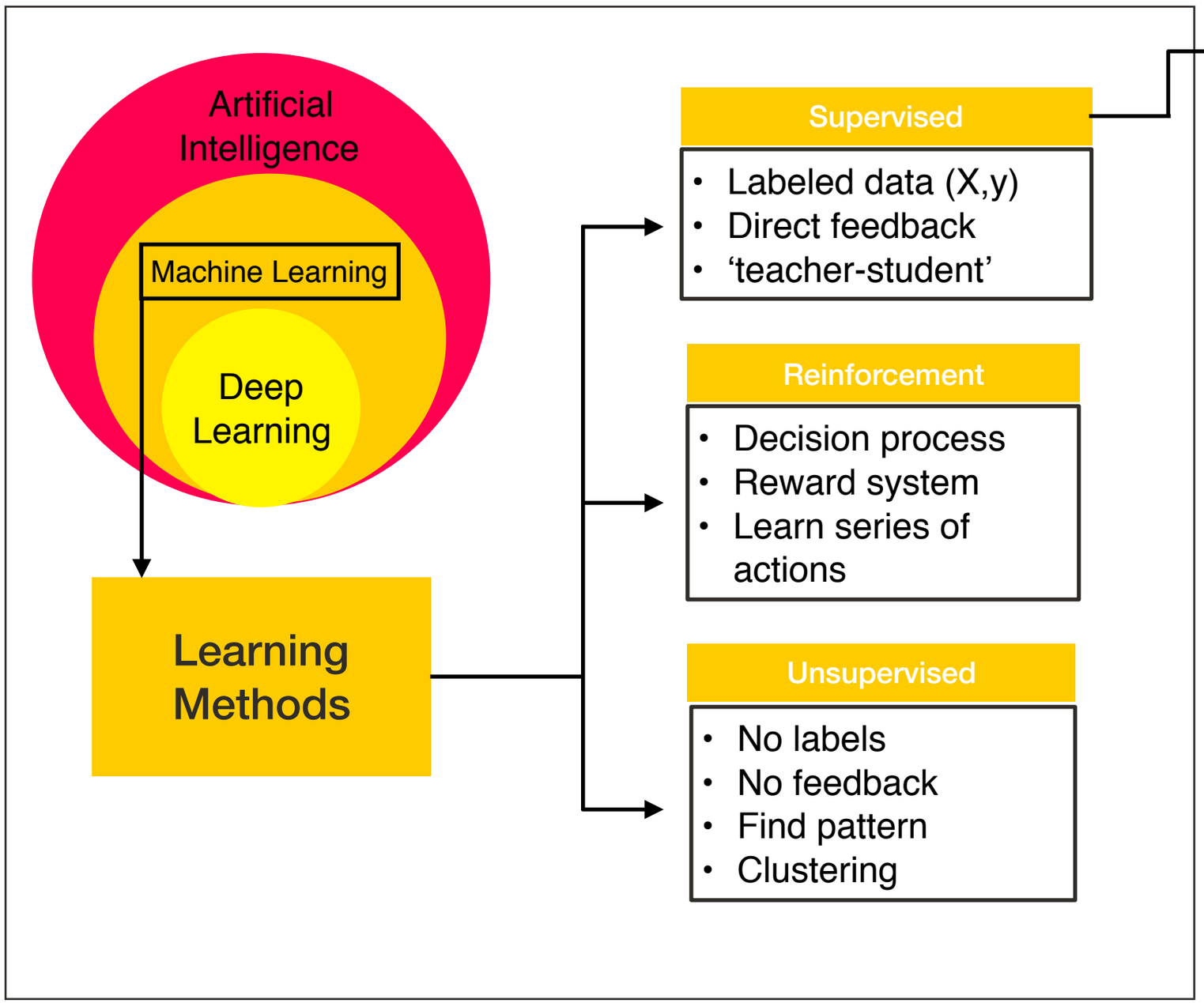


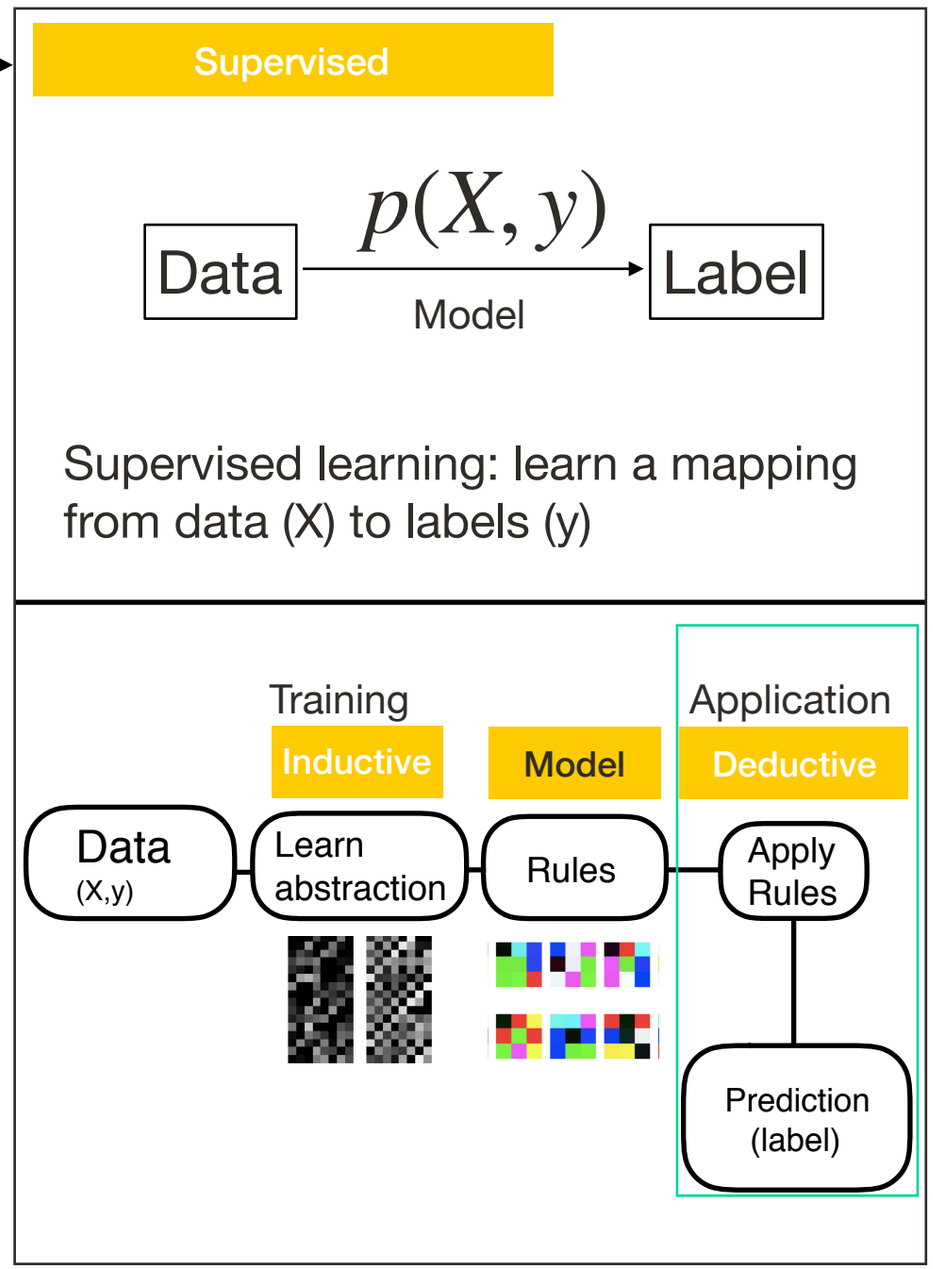
Ben Goertzel



Terminology

- Machine Learning





Terminology

- Bias is an overloaded term

Biased estimator

The difference between the *expected value* and the true value of the parameter being estimated

Bias-variance tradeoff

Bias here is a source of error in your model that causes it to over-generalize and *underfit* your data

Bias and fairness

In machine learning, algorithmic biases are new kinds of bugs. These bugs generically referred as *unwarranted associations*. Such bugs can be harmful to both people and businesses

Ref: Tramer et al

What is a fair decision making model?

A model that respects social and ethical norms, e.g. no discrimination, no racism, no sexism etc.

In practice it is easier to detect unfairness than to prove fairness.

Yet another case

- Houston, we have a problem

TECH

Why it's totally unsurprising that Amazon's recruitment AI was biased against women

BI

Isobel Asher Hamilton
Oct. 13, 2018, 9:00 AM 492

Do your own experiment:
google: professional haircut and observe.
google: unprofessional haircut and observe

The reasons for these problems are rather simple, yet, to find a solution is incredible hard.

BRITISH MEDICAL JOURNAL

LONDON, SATURDAY 5 MARCH 1988

A blot on the profession

Discrimination in medicine against women and members of ethnic minorities has long been suspected,[1-3] but it has now been proved. St George's Hospital Medical School has been found guilty by the Commission for Racial Equality of practising racial and sexual discrimination in its admissions policy.[4] The commission decided not to serve a non-discrimination notice on the school, which it is empowered to do by the Race Relations Act, but as many as 60 applicants each year among 2000 may have been refused an interview purely because of their sex or racial origin. This is a sad finding not only for St George's Hospital Medical School but for the whole profession. It is now important not only that discrimination is swept out of St George's and the profession but also that it is seen to be swept out.

reassuring as it raises the question of what is happening in the other schools.

The commission has made recommendations not just about this particular episode but also about how other schools can avoid similar difficulties. It is emphasised that where a computer program is used as part of the selection process all members of staff taking part have a responsibility to find out what it contains. A major criticism of the staff at St George's was that many had no idea of the contents of the program and those who did failed to report the bias. All staff participating in selection should be trained so that they are aware of the risk of discrimination and try to eliminate it. No one person should have sole responsibility for any stage of the process. The commission recommends that a question on

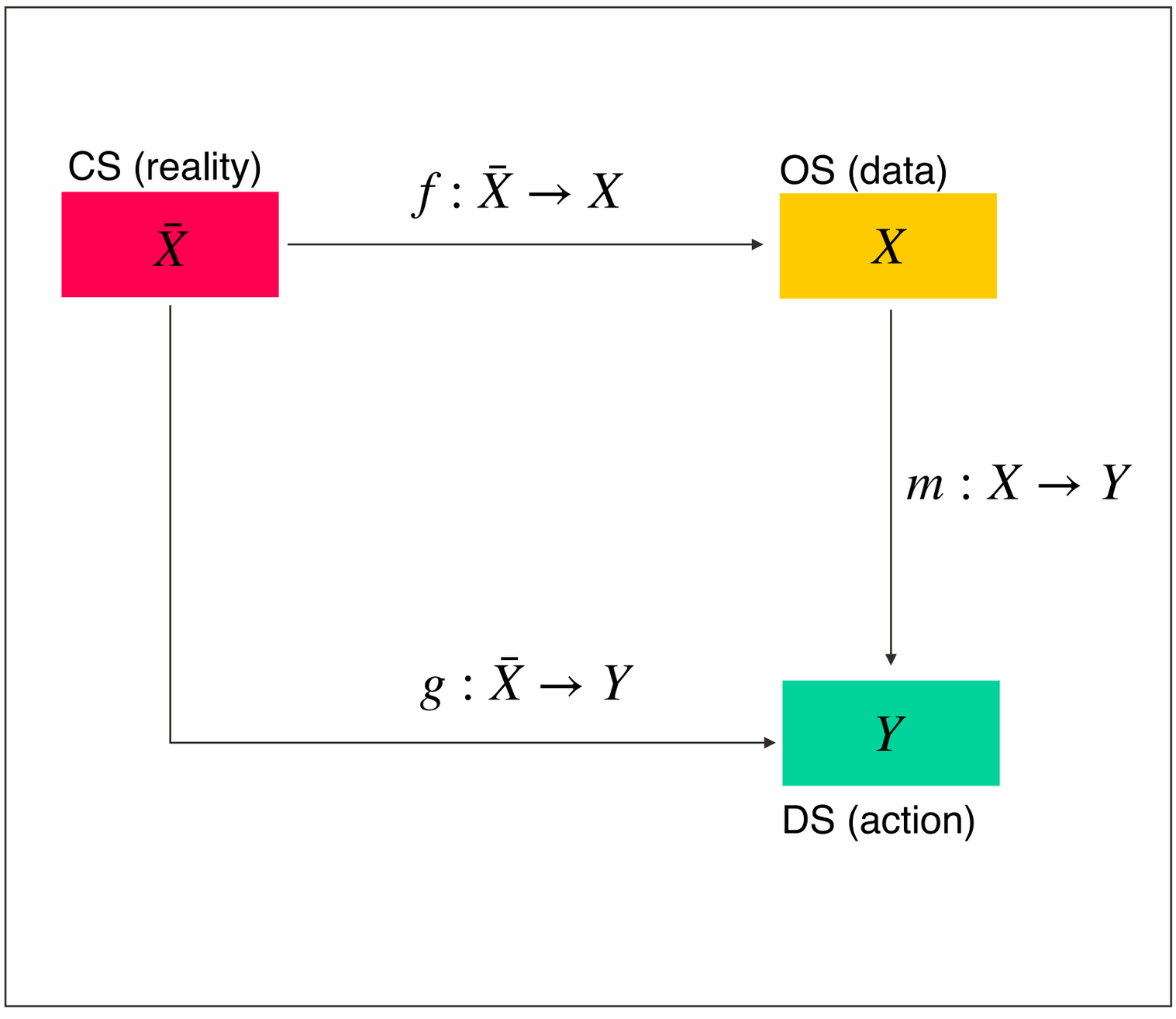
Domains of unfair algorithms

- Unfairness sneaks in everywhere

Jurisprudence	Black people were more likely to be assessed as having a higher risk of recidivism when using commercial prediction tools such as COMPAS
Insurance	An insurance company that used machine learning to workout insurance premiums involuntarily discriminated against elderly patients
Platforms	Google’s image tagger was found to associate racially offensive labels with images of black people

Reality, Data, Decisions

- Sources of biases



f (source of selection bias):
- Considered data and attributes
- Problem to solve with AI/ML

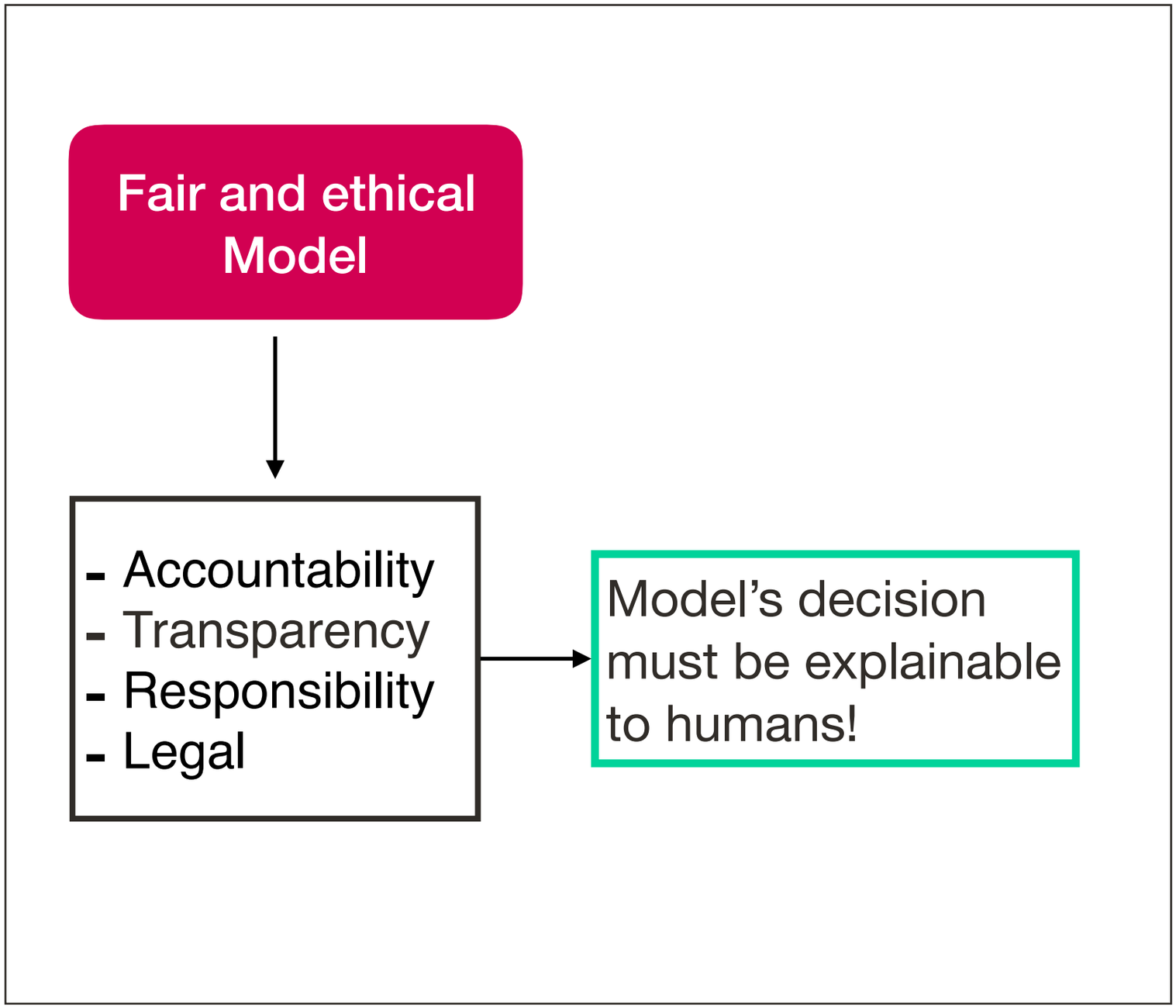
m or model (source of selection and latent bias):
- Model selection
- Training methodologies
- Quality measurement

Y (source of action bias):
- What action do we take?
- Are these actions fair/ethical?

Every step is prone to introduce unfairness.

Model

- What are desirable characteristics of a fair model



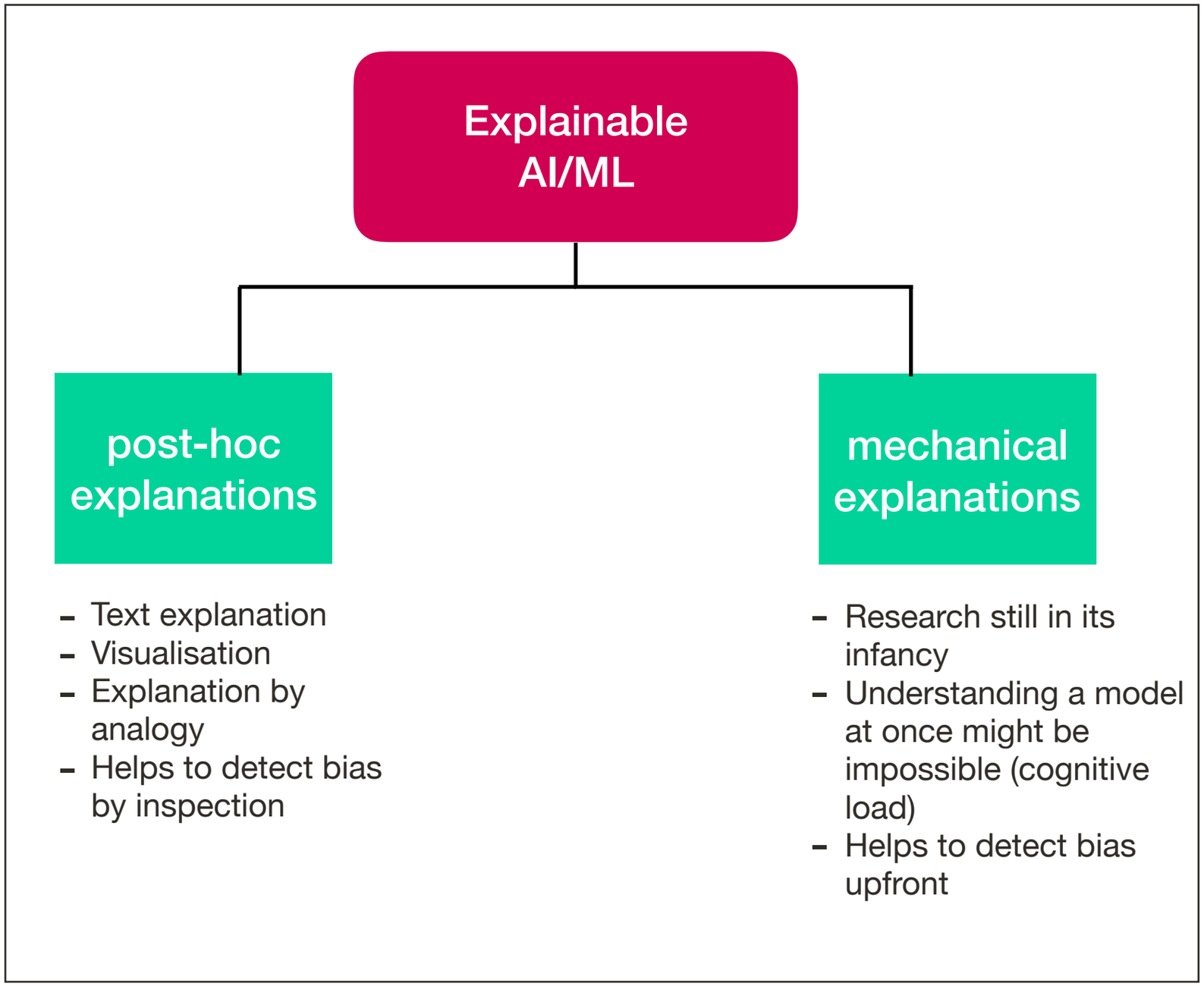
Transparency
- Description and explanation of decision making algorithms' mechanic
- Must be reproducible and understandable by humans

Accountability
- Explain and justify decision
- Must be explained and derived from the decision algorithm
- Must represent moral values and societal norms.

Responsibility
- Refers to algorithms' capability and the role of people interacting with it.
- All stakeholders taking actions based on algorithms' decisions must be accountable

Transparency

- Model transparency by explanations



Post-hoc:

- Gives you the ‘why’ of model decisions
- Disentangles the ‘how’ from the explanation
- Similar to humans’ explanations why a specific decision has been taken

Mechanical:

- Gives you the ‘how’ of model decisions
- Local explanations
- Needs a meta-model of the inner workings of a model

Fair algorithms

- How to detect unfairness?

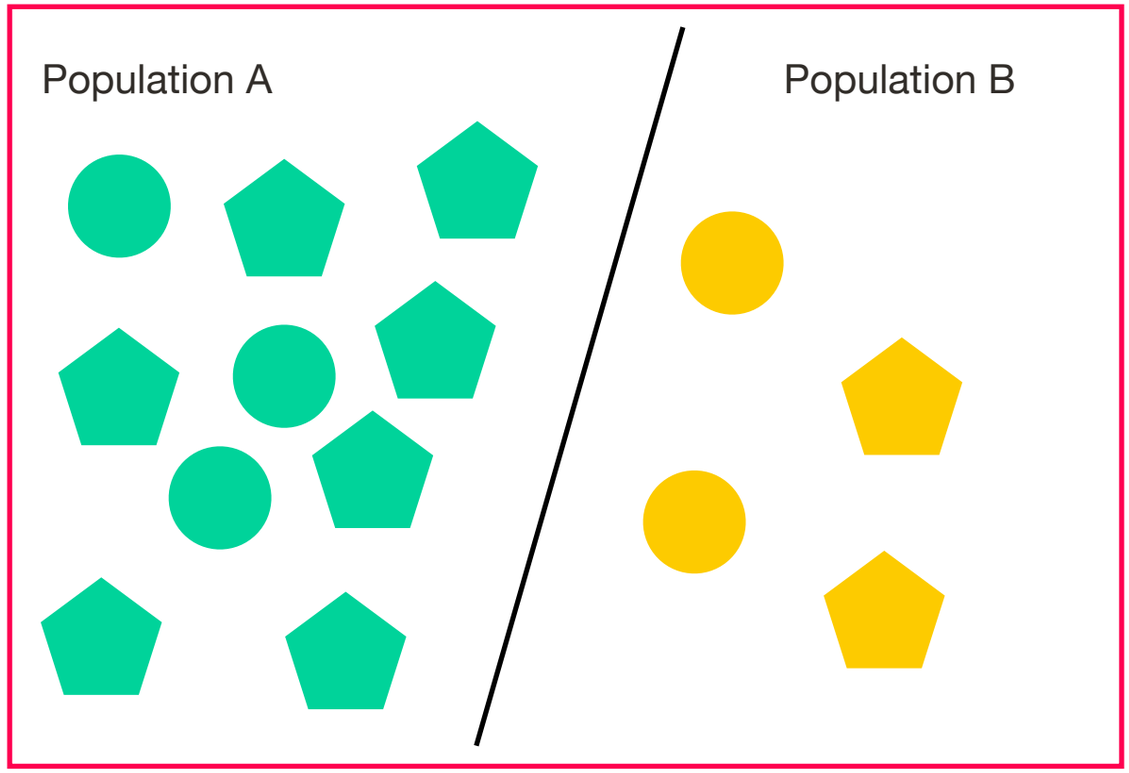
Protected attribute, e.g., male

Protected attribute, e.g., female

Class 1, creditable

Class 1, not creditable

Lets assume we have to solve a two classification problem



Individuals of each population (A and B) share common protected attributes

Data (features) can be divided in protected and unprotected attributes:

$$x = \{x_p, x_u\}$$

Scoring function maps data to a defined interval and decisions depend on defined threshold

$$s(x) \in (s_l, s_u), d(x) = 1 \iff s(x) > t$$

Fair algorithms

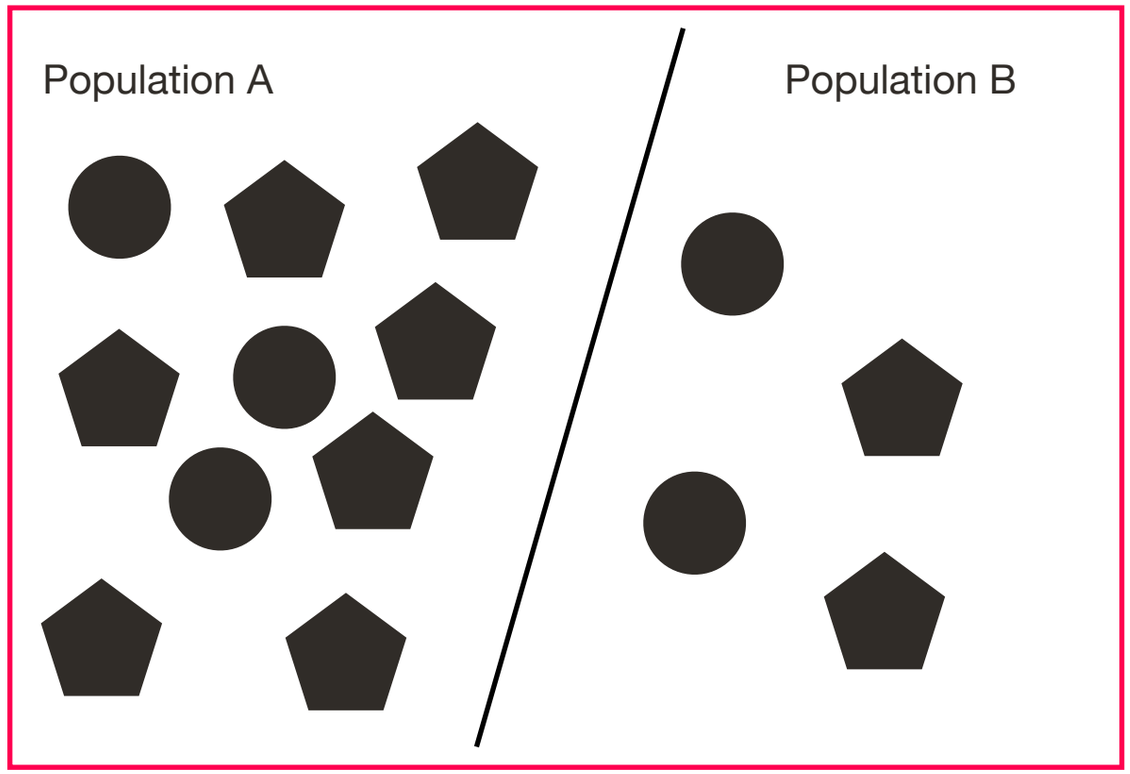
- Anti classification

Fair algorithm:

An algorithm is said to be fair, if the decision function is independent of the protected attributes, e.g., gender, age etc.

Remove protected attributes from data.

Model ‘sees’ unprotected attributes only.



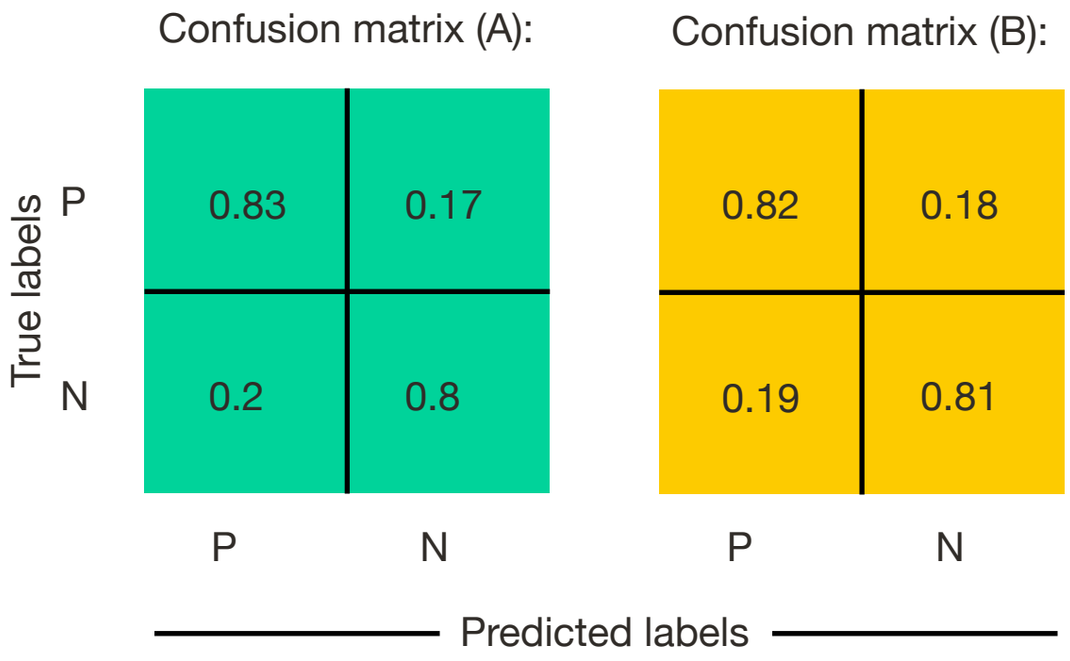
$d(x) = d(x')$ for all x and x', such that $x_u = x'_u$

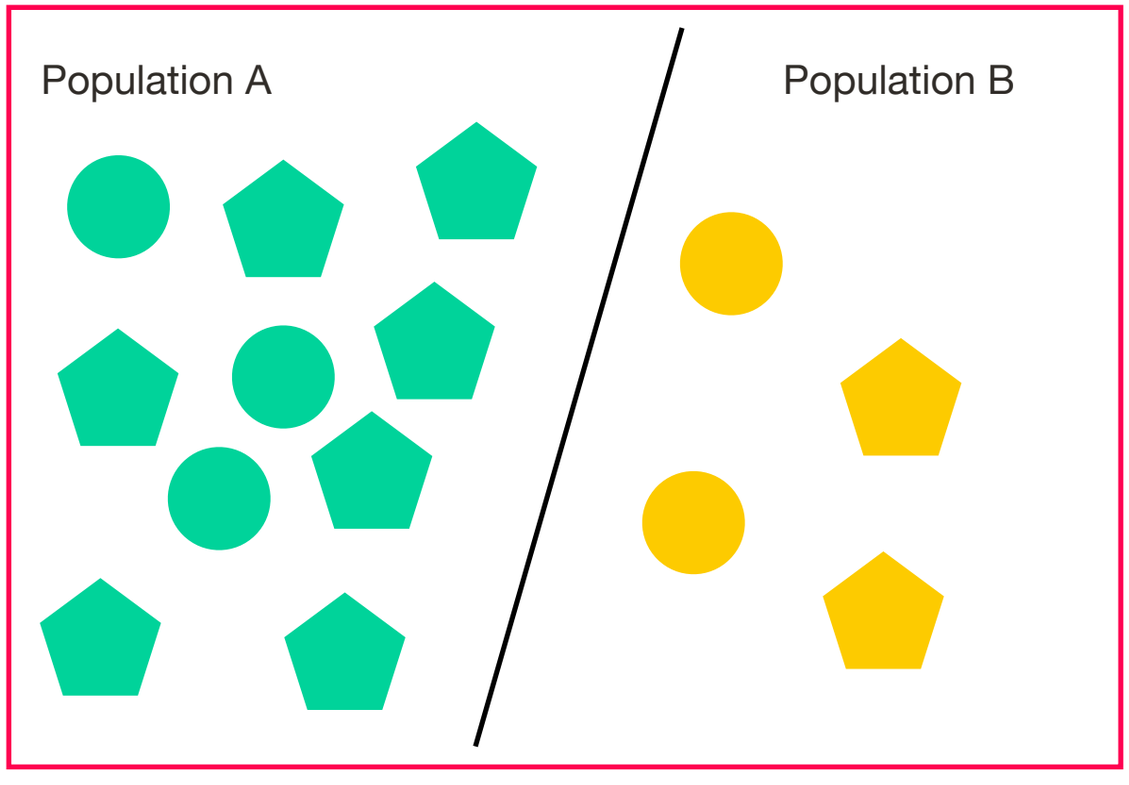
Fair algorithms

- Classification parity

Fair algorithm:

An algorithm is said to be fair, if the classification error is equal across groups defined by the protected attributes,
e.g., gender, age etc.





$$\Pr(d(x) = 1 \mid x_p) = \Pr(d(x) = 1)$$

What to do?

- Challenges

Research

- Definition of fairness
- End-to-end policies to detect and prevent bias and discrimination
- Standardised tests for model quality
- Documentation and knowledge transfer

Industry

- Fairness as a product?
- Know-how
- Fairness vs. productivity
- Be transparent to customers

Politics

- Focus on real problems
- Provide enough funding for research and education
- Think about standards but don’t act prematurely

Individuals

- Ask question
- Make it a topic
- Demand explanations

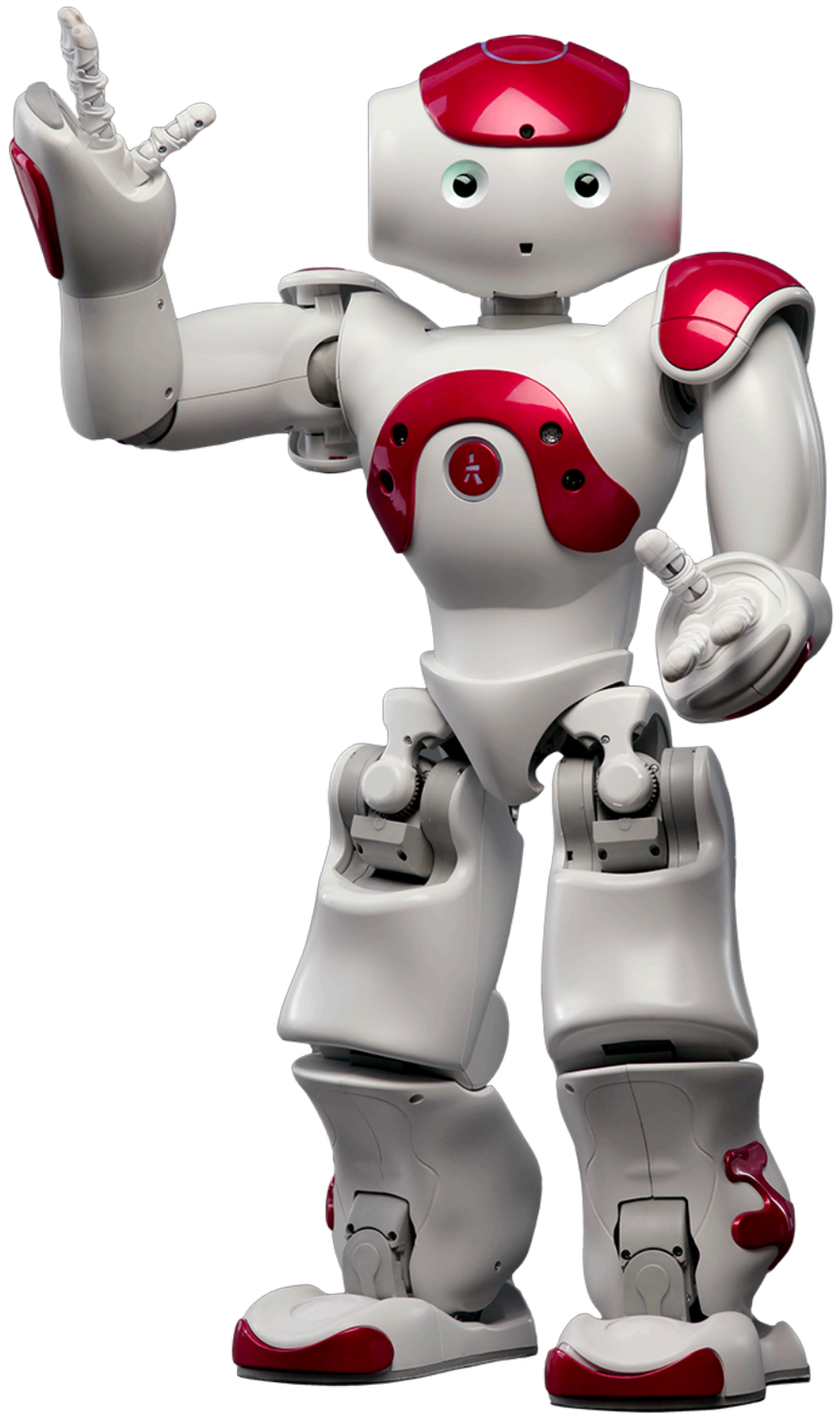
These challenges need an interdisciplinary attitude!

LINK;

https://buchundnetz.com/kieinladung/

Thanks

- That’s the end my friend



Q&A