AI and social justice
Bias and Fairness in AI-based Decision Making

Christoph Heitz
Zurich University of Applied Sciences
Institute of Data Analysis and Process Design
Switzerland

ICTP Workshop on
'Ethical and Societal Challenges of Machine Learning'

Zurich University of Applied Sciences
School of Engineering
IDP Institute of Data Analysis and Process Design

Outlook

- Introduction in ML-based decision making systems?
- The issue of fairness in such systems: why do we have to care for fairness?
- Conceptually: What is fairness?
- Practically: How to measure fairness?
- Concretly: How to build fair ML-based decision systems

Zurich University of Applied Sciences
zh aw School of Engineering
IDP Institute of Data Analysis and Process Design

Ethical and Societal Challenges of AI: The European AI Act

"The AI Act aims to implement an ecosystem of trust by proposing a legal framework within which people use AI-based solutions while encouraging businesses to develop them." (https://www.mondaq.com/india/new-technology/1193996/eu-artificial-intelligence-act-an-overview)

- Draft (2021) under discussion - to be finalized in the next months

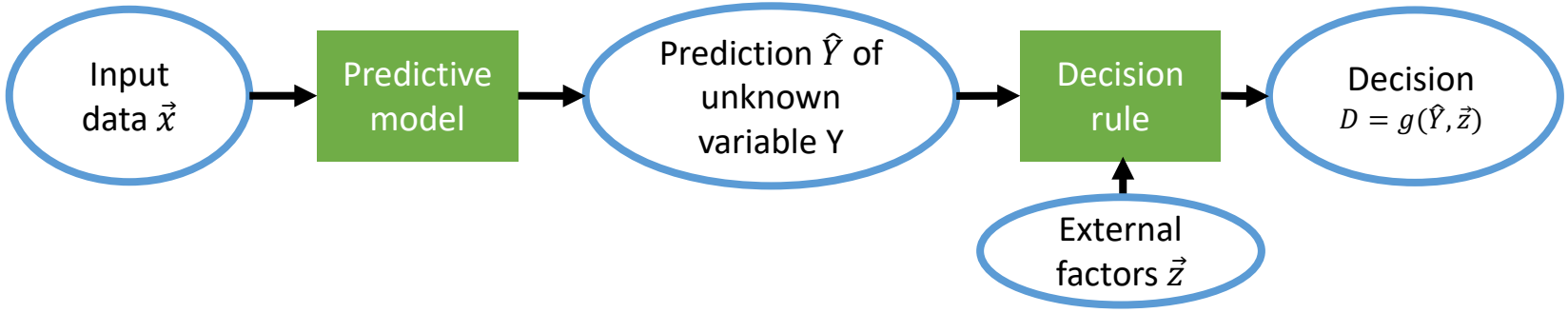
Basis: Ethics guidelines for trustworthy AI (2019)

1. Human Agency and Oversight.
2. Technical Robustness and Safety.
3. Privacy and Data Governance.
4. Transparency.
5. Diversity, Non-discrimination and Fairness.
6. Societal and Environmental Well-being.
7. Accountability.

(see https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf)

Zurich University of Applied Sciences
zhaw School of Engineering
IDP Institute of Data Analysis and Process Design

Data-based decision making



Applications

- Grant loans (banks)
- Individualized insurance premiums
- Algorithmic Hiring
- Predicting Policing
- Law inforcement
- Optimize social care
- Admission to university programs
- …

RETAIL OCTOBER 11, 2018 / 1:04 AM / UPDATED 3 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

Artificial intelligence

Predictive policing algorithms are racist. They need to be dismantled.

Lack of transparency and biased training data mean these tools are not fit for purpose. If we can't fix them, we should ditch them.

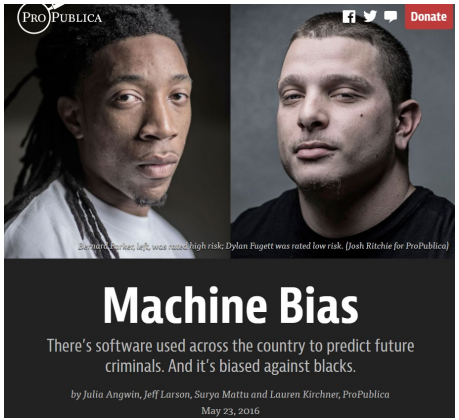
by **Will Douglas Heaven** July 17, 2020

Tech policy / AI Ethics

Can you make an AI that isn't ableist?

IBM researcher Shari Trewin on why bias against disability is much harder to squash than discrimination based on gender or race.

by **Karen Hao** November 28, 2018



IN HER WORDS

Are Algorithms Sexist?

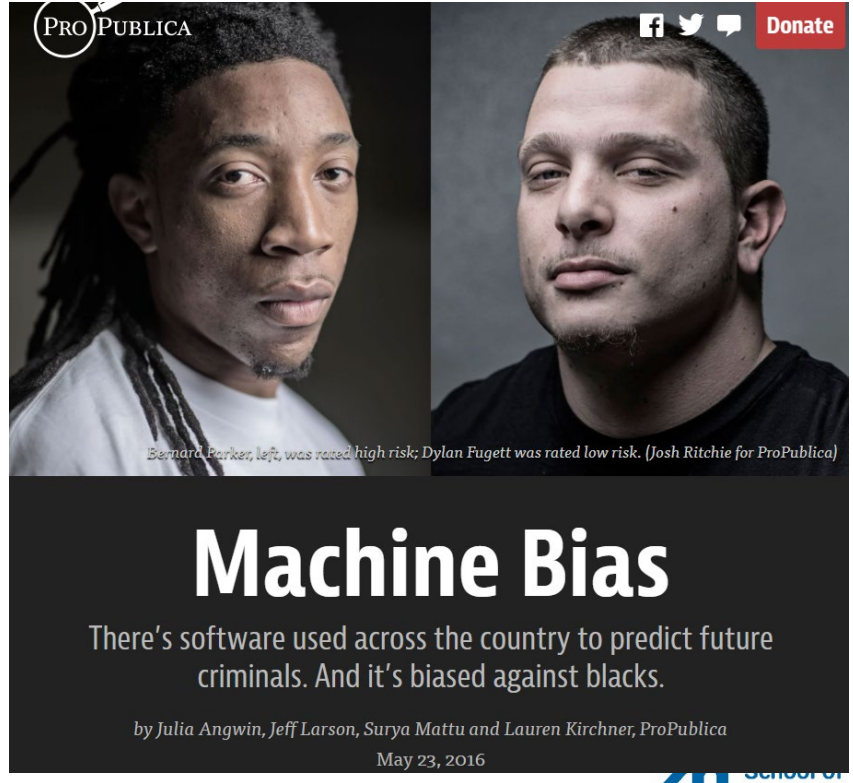
looks around, clears throat Probably, yes.

NEWS · 24 OCTOBER 2019 · UPDATE 26 OCTOBER 2019

Millions of black people affected by racial bias in health-care algorithms

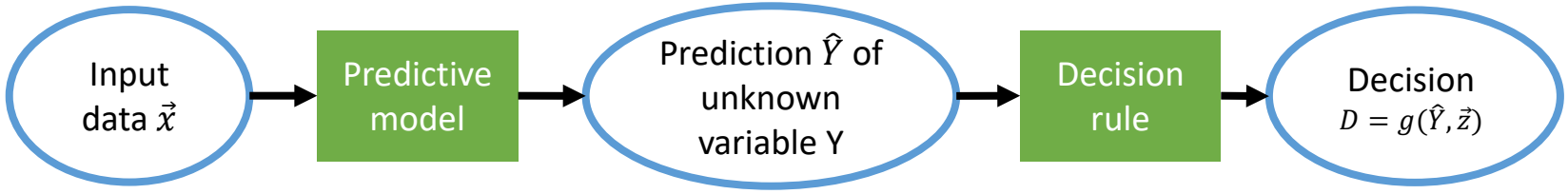
Study reveals rampant racism in decision-making software used by US hospitals — and highlights ways to correct it.

The case of COMPAS



Fairness of decision systems

Fairness is about consequences in people's life



Driver: Goal of decision maker: «making better decision»

- Making more money, saving more lives, …

Consequences: harm/benefit is distributed between groups

- Fairness = moral aspects of this distribution

Example: Bank loan

Loan: 1 M€, for 4 years, interest rate 10%
Customer has a payback probability of p (0<p<1)
Under which conditions should the bank give a loan to the customer?

Solution: p>0.714

Zurich University of Applied Sciences
zh aw School of Engineering
IDP Institute of Data Analysis and Process Design

Which customers to accept? (hypothetical data)

Condition for acceptance:
p>0.714

ID	prediction p
1	0.81
2	0.80
3	0.79
4	0.77
5	0.75
6	0.74
7	0.74
8	0.73
9	0.72
10	0.72
11	0.68
12	0.68
13	0.68
14	0.62
15	0.60
16	0.55
17	0.48
18	0.47
19	0.40
20	0.37

Acceptance threshold

Zurich University of Applied Sciences
School of Engineering
IDP Institute of Data Analysis and Process Design

Looking at gender

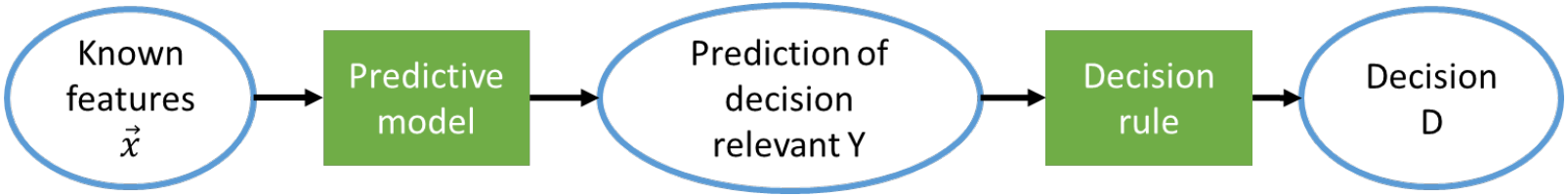
ID	prediction p	decision D	sex
1	0.81	1	m
2	0.80	1	m
3	0.79	1	m
4	0.77	1	m
5	0.75	1	f
6	0.74	1	m
7	0.74	1	m
8	0.73	1	m
9	0.72	1	f
10	0.72	1	f
11	0.68	0	m
12	0.68	0	m
13	0.68	0	f
14	0.62	0	f
15	0.60	0	f
16	0.55	0	m
17	0.48	0	f
18	0.47	0	f
19	0.40	0	f
20	0.37	0	f

Compare men and women:
- What is the chance to receive a loan?

men: 7 out of 10 → 70%
women: 3 out of 10 → 30%

Result:
systematic disadvantage of women

Digging deeper: What is meant with «fairness»?



Most popular definition: **Fairness = nondiscrimination («Group fairness»)**

- We consider social groups, e.g. men and women
- **Does the decision system lead to unjustified inequality? Is one group «worse off» compared to another, in a non-justifiable way?**

Example:
Normative position: «Men and women should have the same chance for a positive decision (D=1)»

- Compare $P(D = 1|m)$ with $P(D = 1|f)$

Fairness would then mean: $P(D = 1|m) = P(D = 1|f)$

What is fairness? - Approaches

Group fairness: Consequences of decisions are equal for groups (on average) ???

Individual fairness: Similar persons receive similar decisions:

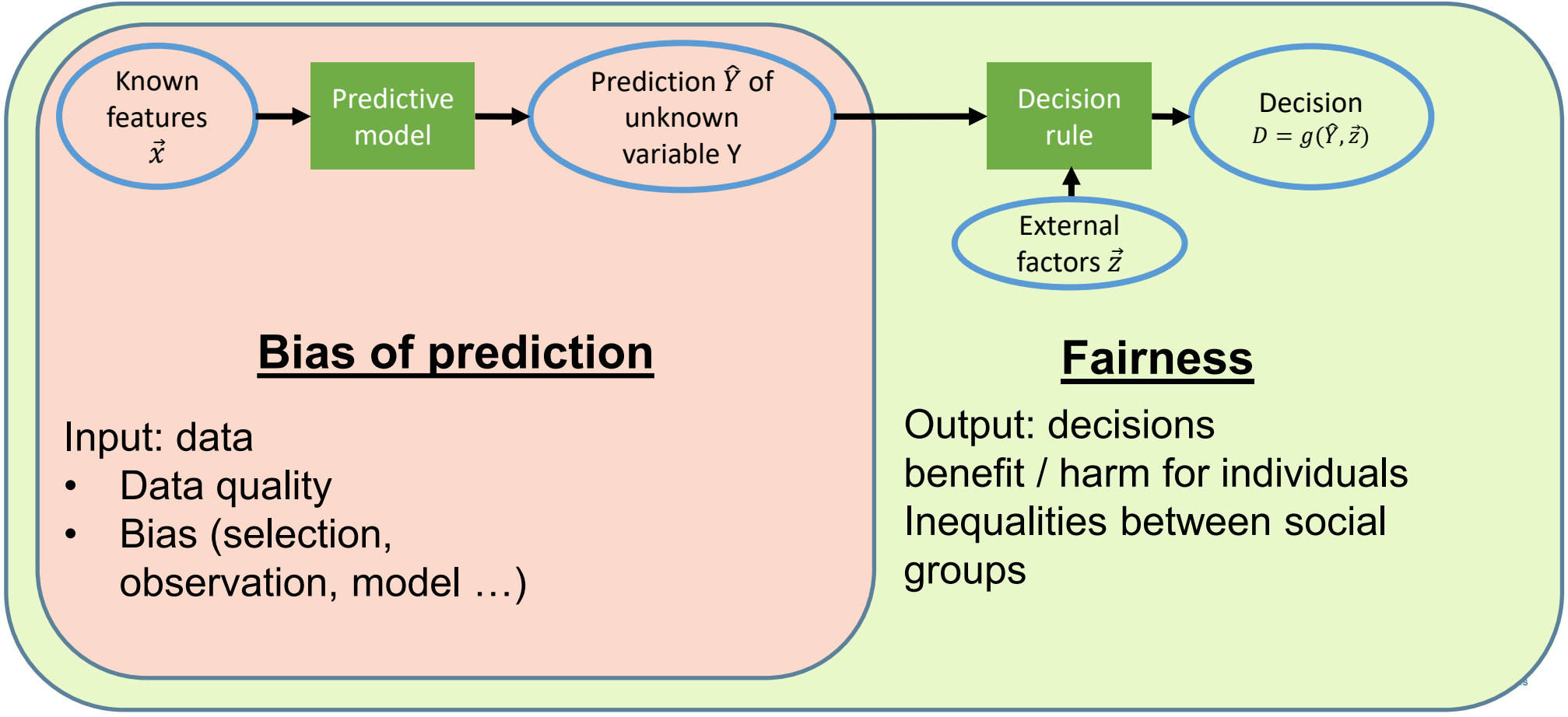
$$\vec{x} \approx \vec{y} \quad \Rightarrow \quad D(\vec{x}) \approx D(\vec{y}) \quad ???$$

Counterfactual fairness: «If I were a woman, the decision is the same!»:

$$\vec{y} = CF(\vec{x}) \quad \Rightarrow \quad D(\vec{x}) = D(\vec{y}) \quad ???$$

Zurich University of Applied Sciences
zh aw School of Engineering
IDP Institute of Data Analysis and Process Design

Sidenote: Bias vs fairness



Different ways of defining fairness

Bank loan:

- Option 1: **Fair is** if men and women applicants have the same chance of acceptance
- Option 2: **Fair is** if, among the customers who are able and willing to pay back, men and women applicants have the same chance of acceptance

Learnings

Fairness might be defined differently!
The appropriate definition depends on **use case** and on **normative position**!

Established Fairness metrics in ML literature

Sources: [20][21][22][23][24][25][26][27] view · talk · edit

		Predicted condition			
	Total population = P + N	**Positive (PP)**	**Negative (PN)**	Informedness, bookmaker informedness (BM) = TPR + TNR − 1	Prevalence threshold (PT) = $\frac{\sqrt{TPR \times FPR} - FPR}{TPR - FPR}$
Actual condition	**Positive (P)**	**True positive (TP),** hit	**False negative (FN),** type II error, miss, underestimation	True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power = $\frac{TP}{P}$ = 1 − FNR	False negative rate (FNR), miss rate = $\frac{FN}{P}$ = 1 − TPR
	Negative (N)	**False positive (FP),** type I error, false alarm, overestimation	**True negative (TN),** correct rejection	False positive rate (FPR), probability of false alarm, fall-out = $\frac{FP}{N}$ = 1 − TNR	True negative rate (TNR), specificity (SPC), selectivity = $\frac{TN}{N}$ = 1 − FPR
	Prevalence = $\frac{P}{P+N}$	Positive predictive value (PPV), precision = $\frac{TP}{PP}$ = 1 − FDR	False omission rate (FOR) = $\frac{FN}{PN}$ = 1 − NPV	Positive likelihood ratio (LR+) = $\frac{TPR}{FPR}$	Negative likelihood ratio (LR−) = $\frac{FNR}{TNR}$
	Accuracy (ACC) = $\frac{TP+TN}{P+N}$	False discovery rate (FDR) = $\frac{FP}{PP}$ = 1 − PPV	Negative predictive value (NPV) = $\frac{TN}{PN}$ = 1 − FOR	Markedness (MK), deltaP (Δp) = PPV + NPV − 1	Diagnostic odds ratio (DOR) = $\frac{LR+}{LR-}$
	Balanced accuracy (BA) = $\frac{TPR+TNR}{2}$	F_1 score = $\frac{2PPV \times TPR}{PPV+TPR} = \frac{2TP}{2TP+FP+FN}$	Fowlkes–Mallows index (FM) = $\sqrt{PPV \times TPR}$	Matthews correlation coefficient (MCC) = $\sqrt{TPR \times TNR \times PPV \times NPV} - \sqrt{FNR \times FPR \times FOR \times FDR}$	Threat score (TS), critical success index (CSI), Jaccard index = $\frac{TP}{TP+FN+FP}$

Zurich University of Applied Sciences
zhaw School of Engineering
IDP Institute of Data Analysis and Process Design

Two problems

1. **Impossibility theorems**: It is not possible for a decision system to be fair with respect to all fairness metrics

It is even worse: The different metrics exclude each other

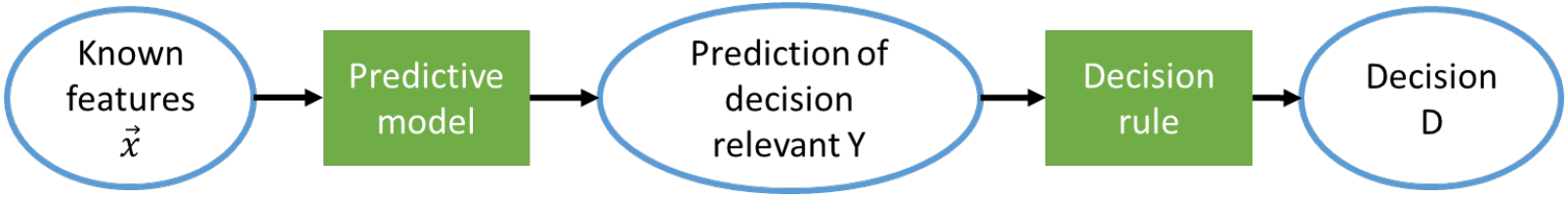
- If a system is fair according to one metric, it is unfair with respect to most others

2. **How to find the «appropriate» fairness definition?**

This is an ethical questions, not a technical one!

- «Techies» cannot decide this (alone) – responsibility is with the owner of the system
- BUT: «Technies» have to be aware of the problem and ask for a decision

Reality check: Unfairness is the rule, not the exception!



Maximization of decision makers utility does not care about fairness!

It is pure luck if a prediction-based decision system is fair!

- In most cases, it is not, unless fairness is explicitly built in!

Many real-world examples show this.

How to build a fair ML-based decision system?

Each decision system results in

- Degree of goal achievement of decision maker
  - Average over the full population of decision subjects
- Degree of fairness
  - Based on some fairness metric

Task: Maximize goal achievement, while still caring for fairness

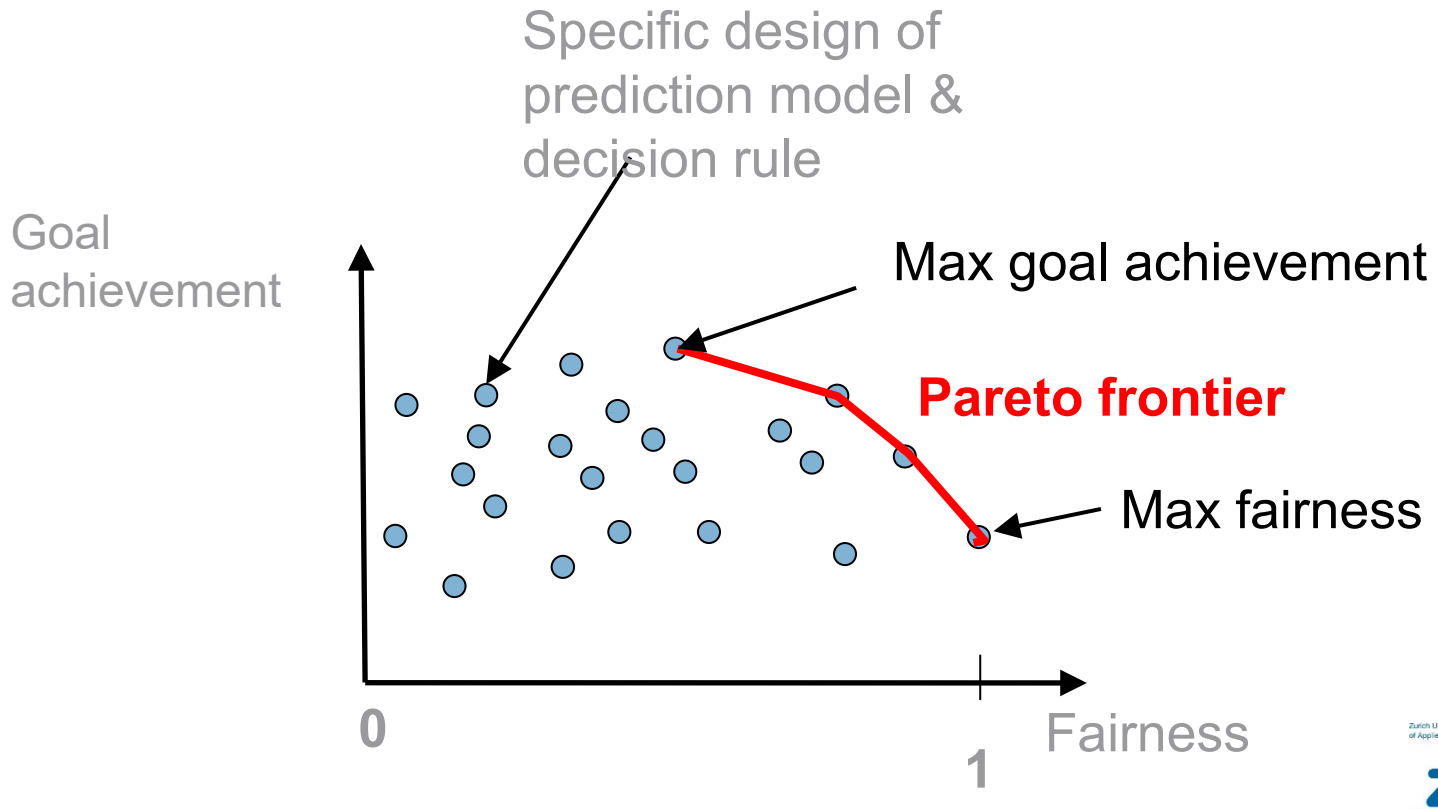
Approaches:

- Solve constraint optimization problem
  - Maximize goal, with respect to «degree of fairness > F_0»)

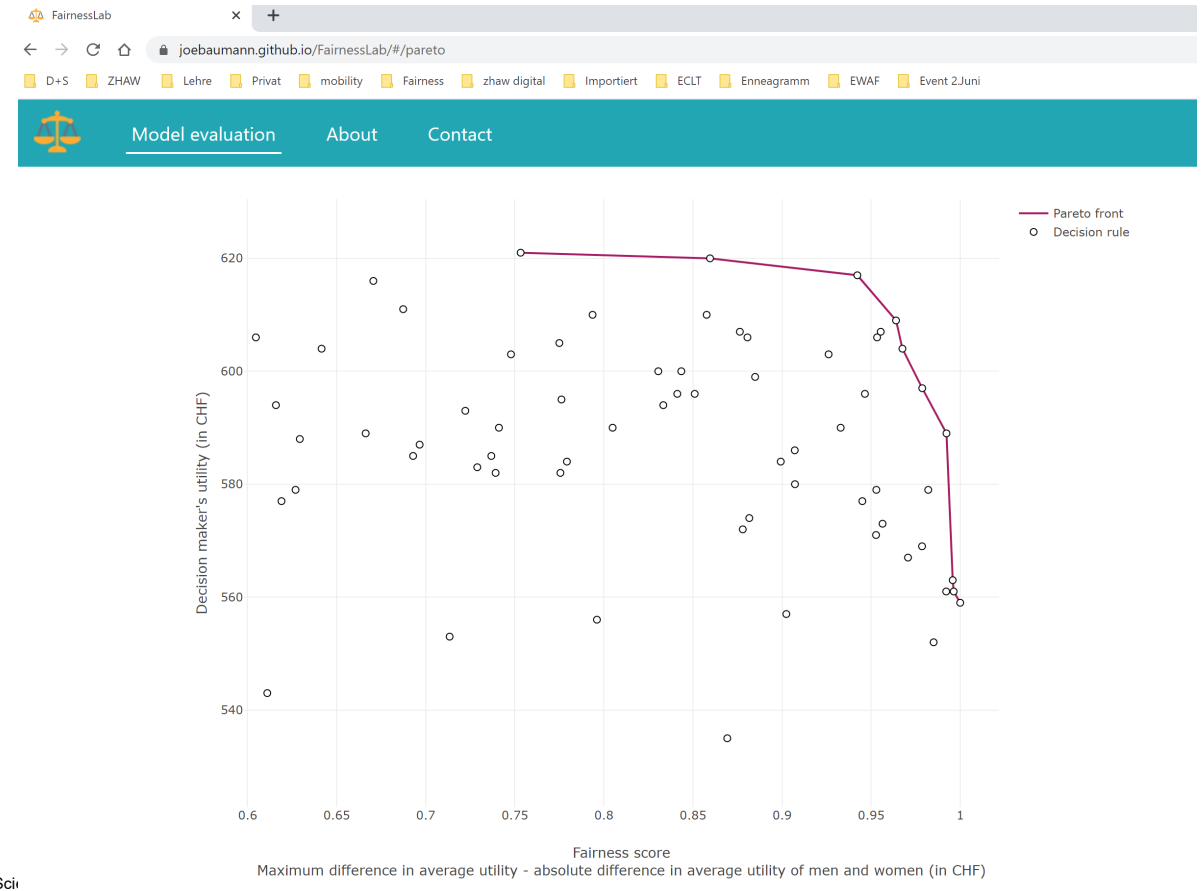
- Analyze different designs of decision system with respect to these two variables (goal achievement, fairness) and find optimum combination
  - Multicriteria optimization

Zurich University of Applied Sciences
zh aw School of Engineering
IDP Institute of Data Analysis and Process Design

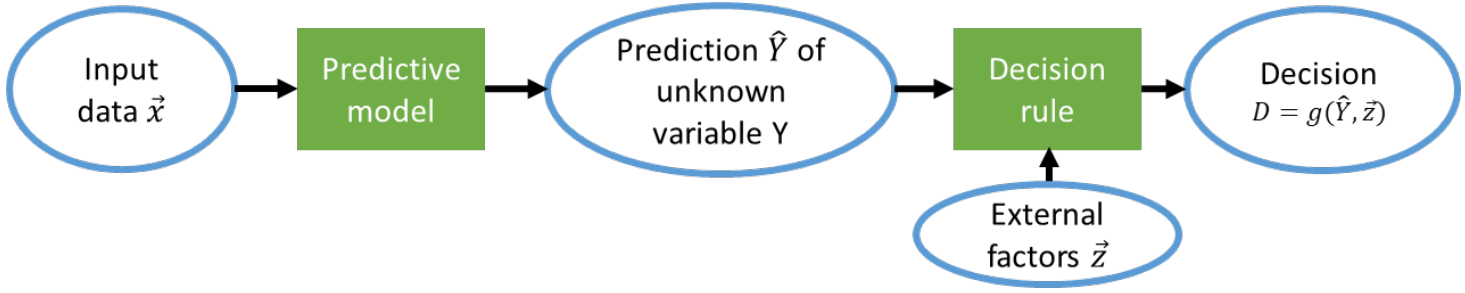
Goal achievement and Fairness – the cost of fairness



The fairness lab: trade-off between fairness and goal achievement



Technical solutions for achieving fairness



Preprocessing: Find better predictive model by manipulating learning data

In-Processing: Find better predictive model by manipulating learning strategy

Post-Processing: Take predictive model as it is, and find better decision rule

Conclusion

Applying ML for «making better decisions» normally leads to social injustice (unfairness), unless fairness is explicitly built in

Fairness can be measured (fairness metrics) – different definitions possible

- They reflect what «fairness» means in the specific context
- A choice has to be made (moral analysis)

Fairness can be implemented technically

- Two dimensions to be distinguished: Goal achievement / fairness
- Task: find optimum solutions
- Methods: pre-processing/in-processing/post-processing

Some references

Barocas, S., Hardt, M., and Narayanan, A. (2019). *Fairness and Machine Learning*. fairmlbook.org.

Michael Kearns and Aaron Roth. The Ethical Algorithm: The Science of Socially Aware Algorithm Design. Oxford University Press, Inc., USA, 2019.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. ProPublica, May, 23(2016):139–159, 2016.

Chouldechova, A. (2017). Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big data*, 5(2):153–163.

Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. The (Im)possibility of fairness. Communications of the ACM, 64(4):136–143, 4 2021.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of Opportunity in Supervised Learning. In Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16, page 3323–3331, Red Hook, NY, USA,2016. Curran Associates Inc.

Hertweck, C., Baumann, J., Loi, M., Viganò, E., and Heitz, C. (June 2022): A Justice-Based Framework for the Analysis of Algorithmic Fairness-Utility Trade-Offs, http://arxiv.org/abs/2206.02891

Baumann, J., Hertweck, C., Loi, M., and Heitz, C. (June 2022): Distributive Justice as the Foundational Premise of Fair ML: Unification, Extension, and Interpretation of Group Fairness Metrics. http://arxiv.org/abs/2206.02897

Thank you for your attention!

Happy to answer any questions!

Christoph.heitz@zhaw.ch