Algorithmic Fairness
A Major Challenge Area for Ethics of Data-Based Business

Christoph Heitz
ZHAW Zurich University of Applied Sciences
Digital Society Initiative, University of Zurich

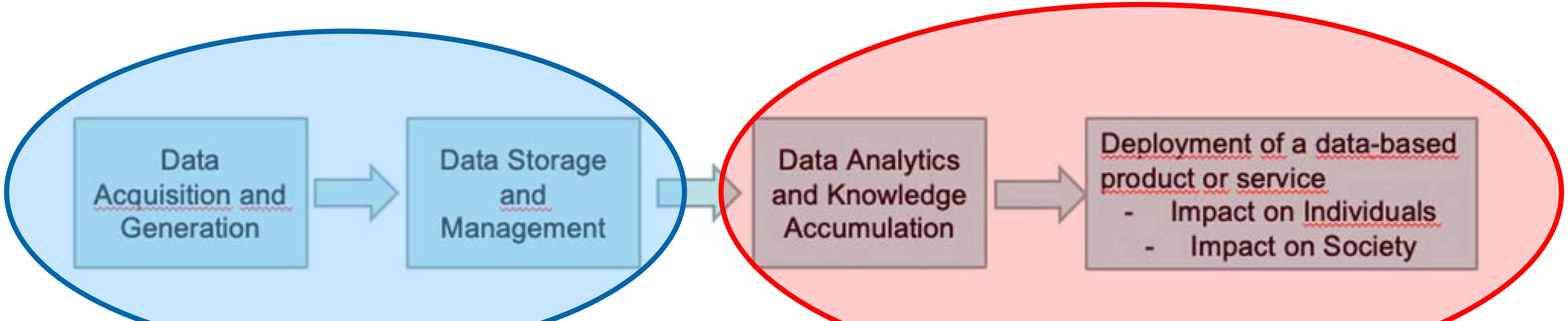
in cooperation with Michele Loi and Markus Christen, DSI, Univ. of Zurich

Zürcher Hochschule für Angewandte Wissenschaften
zh aw School of Engineering

Universität Zürich[UZH]
Digital Society Initiative

data service
Swiss Alliance for Data-Intensive Services

The data science pipeline and ethics



Loi, M., C. Heitz, A. Ferrario, A. Schmid, and M. Christen. 2019. "Towards an Ethical Code for Data-Based Business."

Ethical issues

Data Privacy
Data Protection

Impact on our world?
Threat of societal values, e.g.
- Freedom
- Justice and fairness
- …

The COMPAS Case

- 2016: ProPublica investigates a risk assessment tool for criminal recidivism (COMPAS)
  - developed by a private company (Northpointe)
  - used in many US states over years (>1 Mio criminals assessed)
- ProPublica showed that the tool was racially biased
  - black people more likely to be wrongly predicted to re-offend than white people
- Northpointe had to change its name (now equivant) as a consequence of the public debate

Julia Angwin, Jeff Larson. 2016. "Machine Bias." Text/html. ProPublica. May 23, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Amazon‘s sexist hiring algorithm

amazon

- 2014: Amazon starts building algorithms to review job applicants
- 2015: Amazon detects gender bias for software developer jobs
  - Reason: male-specific expressions
- Attempts to remove gender bias failed (!)
- 2017: Amazon announces the stop of the program, trying to limit image problems

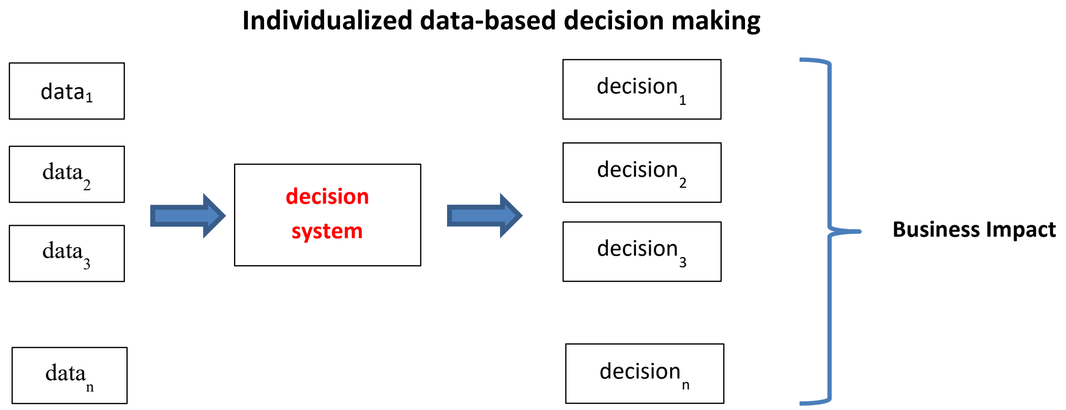
Reuters. 2018. “Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women,” October 10, 2018. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.

The Austrian AMS

- 2018: The Austrian Public Employment Service Austria (Arbeitsmarktservice AMS) announces the introduction of a software sorting unemployed people according to their chances on the job market.
- Prediction model developed by private company Synthesis GmbH
- Prediction uses a regression model
  - Factor “female” has a negative coefficient (Der Standard, 20.10.2018)
- Public debate about efficiency vs. fairness – still ongoing

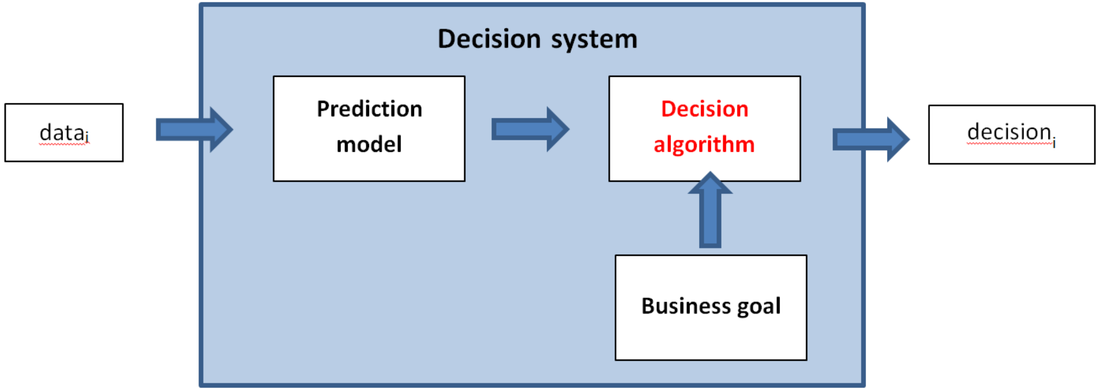
Holl, Jürgen, Günter Kernbeiß, and Michael Wagner-Pinter. 2018. “Das AMS-Arbeitsmarktchancen-Modell,”

Context: Data-based decisions in business



- ☐ Individualized decision making on humans, based on their data

Typical ML case: decision based on prediction



- Goal: maximize business goal by taking individualized decisions, based on prediction
  - E.g. credit risk, risk of recidivism, risk of failing, …
- Driver: Huge business potential to be harvested

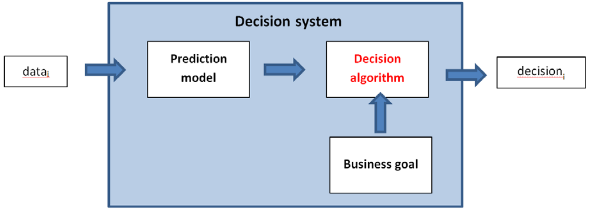
Algorithmic bias in data-based decisions?

Zürcher Hochschule
für Angewandte Wissenschaften
School of Engineering

Definition «algorithmic bias» (https://en.wikipedia.org/wiki/Algorithmic_bias):

- ☐ Algorithm: sets of instructions within computer programs that determine how these programs read, collect, process, and analyze data to generate some readable form of analysis or output.
- ☐ The term *algorithmic bias* describes **systematic and repeatable errors that create unfair outcomes**, such as privileging one arbitrary group of users over others.

- ☐ **Problem 1:** Data-based decision algorithms are typically biased
  - › Business goal optimization does not care about bias!
- ☐ **Problem 2:** Developers do not care
  - › Many are not even aware of the problem of bias
- ☐ **Problem 3:** Unfair algorithms are actually implemented
  - › Reputation risk, negative societal impact



Algorithmic bias in research

Zürcher Hochschule für Angewandte Wissenschaften
zh aw School of Engineering

- Issue is on the research agenda since about 2015
- Many publications in the Machine Learning community
  - Reasons for bias (inappropriate data, suboptimal learning procedures, algorithmic issues, ….)
  - Important result: just ignoring sensitive variables („Fairness Through Unawareness») does not do the job
  - Countermeasures for different prediction algorithms developed
  - Etc.
- Conceptual learnings
  - Fairness can be measured by statistical properties of prediction or decision algorithm
  - Fairness can be defined in different ways

COMPAS revisited

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

- For binary prediction problems: Confusion matrix
- COMPAS: 1 = re-offend, 0 = not re-offend
- Result: FP rate higher for black people → „unfair“

Actual Values

Predicted Values	Positive (1)	Negative (0)
Positive (1)	TP	FP
Negative (0)	FN	TN

Fairness criteria

Zürcher Hochschule für Angewandte Wissenschaften
School of Engineering

- Simplest problem statement:
  - Consider two groups (A and B)
  - Consider a prediction of binary variable Y: prediction = $\hat{y}$, true value=y
  - Decision = prediction: $D = \hat{y}$
- Some fairness criteria:
  - Demographic parity: $P\left[D=1 \middle| A\right] = P\left[D=1 \middle| B\right]$
  - Equal FPR $P\left[D=1 \middle| y=0, A\right] = P\left[D=1 \middle| y=0, B\right]$
  - Equal odds = Equal FPR and Equal TNR
  - Equal Positive Predictive Value: $P\left[y=1 \middle| \hat{y}=1, A\right] = P\left[y=1 \middle| \hat{y}=1, B\right]$

What is fair? - Fairness definitions

	Definition	Paper	Citation #	Result
3.1.1	Group fairness or statistical parity	[12]	208	×
3.1.2	Conditional statistical parity	[11]	29	✓
3.2.1	Predictive parity	[10]	57	✓
3.2.2	False positive error rate balance	[10]	57	×
3.2.3	False negative error rate balance	[10]	57	✓
3.2.4	Equalised odds	[14]	106	×
3.2.5	Conditional use accuracy equality	[8]	18	×
3.2.6	Overall accuracy equality	[8]	18	✓
3.2.7	Treatment equality	[8]	18	×
3.3.1	Test-fairness or calibration	[10]	57	✓✗
3.3.2	Well calibration	[16]	81	✓✗
3.3.3	Balance for positive class	[16]	81	✓
3.3.4	Balance for negative class	[16]	81	×
4.1	Causal discrimination	[13]	1	×
4.2	Fairness through unawareness	[17]	14	✓
4.3	Fairness through awareness	[12]	208	×
5.1	Counterfactual fairness	[17]	14	–
5.2	No unresolved discrimination	[15]	14	–
5.3	No proxy discrimination	[15]	14	–
5.4	Fair inference	[19]	6	–

Table 1: Considered Definitions of Fairness

Verma, Sahil, and Julia Rubin. 2018. “Fairness Definitions Explained.” In Proceedings of the International Workshop on Software Fairness - FairWare ’18, 1–7. Gothenburg, Sweden: ACM Press. https://doi.org/10.1145/3194770.3194776

- Fairness can be defined differently
  - E.g. Arvind Narayanan (FAT* 2018): Tutorial: 21 fairness definitions and their politics
- Typically, <u>different fairness criteria are mutually exclusive</u>: They cannot be met simultaneously! (Kleinberg et al 2016)
- **A choice has to be made!**

COMPAS revisited (II)

Zürcher Hochschule für Angewandte Wissenschaften
School of Engineering

- COMPAS actually fulfills an important fairness criterion: positive predictive value (PPV) is well met (Kleinberg et al 2016, Chouldechova 2017)
- But: FPR and FNR are different for blacks and whites → this was what ProPublica brought up
- It can be shown for arbitrary prediction algorithms (Chouldechova 2017):

$$FPR = \underbrace{\frac{p}{1-p}}_{\text{prevalence}} \frac{1-PPV}{PPV}(1-FNR)$$

No prediction algorithm can meet both fairness citeria simultaneously!

What is fair?

- Fairness and justice has a long history in moral and political philosophy
- Equal rules for all (procedural fairness)
  - Business potential lies exactly in discrimination!
- So we have to analyse the consequences
  - Consequentialist ethics
- Different philosophical concepts of fairness and justice, e.g.
  - Welfare economics and utilitarism
  - different theories to explain what makes discrimination wrong

The problem of algorithmic fairness

Zürcher Hochschule für Angewandte Wissenschaften
School of Engineering

For developing a „fair algorithm“, two problems have to be solved

- An ethical choice problem (decision): What is fair?
  - may depend on the concrete situation
  - Is an ethical question, not a technical one
  - choice must be justified and defended (towards customers and society)
  - Result: fairness criterion expressed in statistical terms (measurable)
- A technical problem: Create a decision algorithm that meets the specified fairness criterion
  - ML literature shows some solutions for some fairness criteria, but not a general solution procedure
  - Issues: Input data for learning procedures? How to train models? How to assess decision models? …

Necessary: Integration of ethics and engineering!

Integrated solution approach

Ethical decision making: What is fair? → (Statistical fairness criterion defined) → Technical implementation → (Defined fairness criterion met) → „Fair“ algorithm

Ethical decision making: What is fair?

- Based on solid philosophical concepts
- Structured approach (discourse)
- Do-able for non-philosophers (managers and Data Scientists!)

Technical implementation

- Maximization of business goal with fairness constraints, or
- Multicriteria optimization
- „Fairness by design“

„Fair“ algorithm

- Assessment possible

Zürcher Hochschule
für Angewandte Wissenschaften

zh
aw

School of Engineering

Conclusion

- ☐ Algorithmic fairness is an important issue for all companies doing data-based business
  - › Second big issue after data privacy and protection
  - › Ethical responsability AND economic risk
- ☐ Fairness is an ethical issue, not primarily a technical one
  - › Different fairness definitions possible
  - › What is considered fair depends on situation and stakeholders
- ☐ Creating fair algorithms needs the <u>combination of an ethical decision making process</u> (which fairness do we want to produce?) <u>with a technical solution method</u> (how to produce this fairness?)
  - › Ethical discourse needs integration of all stakeholders - engineering can‘t do it alone!
  - › Specific expertise is needed for the model builders – often a problem today
- ☐ Field is new, up to now no integrated methodology is available to make sure that decision algorithms are fair in a well-defined, understood and explainable way
  - › There is some work to do!

Thank you for your attention!