(un)Fairness in algorithmic decision-making

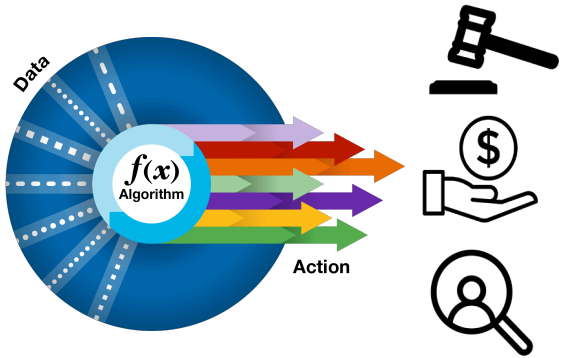
Isabel Valera

Saarland University & MPI for Software Systems

"I'm right there in the room, and no one even acknowledges me."

Gender Summit, 15th April 2021

Machine learning to inform consequential decisions



Automated data analysis supplements and even replaces human supervision in consequential decision-making

How do we make sure that they are aligned with human goals & values?

"Which features about applicants should I collect to maximize profit? "

"Is the system fulfilling existing discrimination laws?"

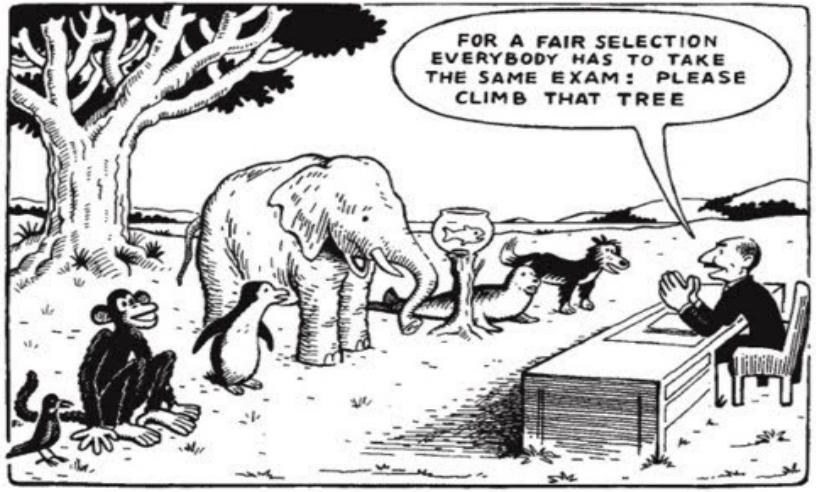
"Why I did I get this outcome; what should I do to get a more favourable outcome next time?"

Machine Learning can be unfair to many

WHAT is (un)Fairness in ML?

Discrimination

[...] **wrongfully** impose a **relative disadvantage** on persons based on their membership in some **salient social group,**
e.g., race or gender. [Altam'16]



In ML, each individual as a set of features

$(\boldsymbol{x}, z)$

Non-sensitive features

Sensitive feature

HOW is (un)Fairness in ML?

Historical data:

	Sensitive features	Non-sensitive features			Outcomes
		Income	Credit card	Previous loans	Repaid?
Applicant 1		1.5k	VISA	✓	
Applicant 2		2.7k	VISA	✗	

New applicant:

New Applicant		2.2k	VISA	✓	?

HOW is (un)Fairness in ML?

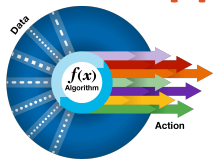
Historical data:

	Sensitive features	Non-sensitive features			Outcomes
		Income	Credit card	Previous loans	Repaid?
Applicant 1		1.5k	VISA	✓	PAID IN FULL
Applicant 2		2.7k	VISA	✗	default?

New applicant:

New Applicant		2.2k	VISA	✓

Prediction!



PAID IN FULL

WHY (un)Fairness in ML?

Features may be less informative or reliably collected for minority group(s)

Less data from minority groups lead to higher errors

Each person as a set of features

$(\boldsymbol{x}, z)$

Non-sensitive features

Sensitive feature

Historical data contains human biases and stereotypes

AI prone to finding proxies for sensitive information

Decisions lead to skewed data leading to poor predictions

Advances on Fair ML

Fairness in ML

Challenge #1: From definitions to measures

How to measure fairness?

- Formalizing or interpreting a fuzzy definition to make it measurable for empirical observations
- Trade-off different measurements of fairness

Challenge #2: Mechanisms

How to incorporate fairness into algorithmic decision making?

- Understand (un)fairness in AI
- Mechanisms to ensure fairness
- Trade-off fairness and accuracy

Fairness in ML - Definitions

Z Sensitive feature

Y Target (label)

$\hat{Y}$ Prediction

Independence: Require the prediction and the sensitive features to be independent:

$$\hat{Y} \perp Z$$

Separation: Require the prediction and the sensitive features to be independent conditional on the target:

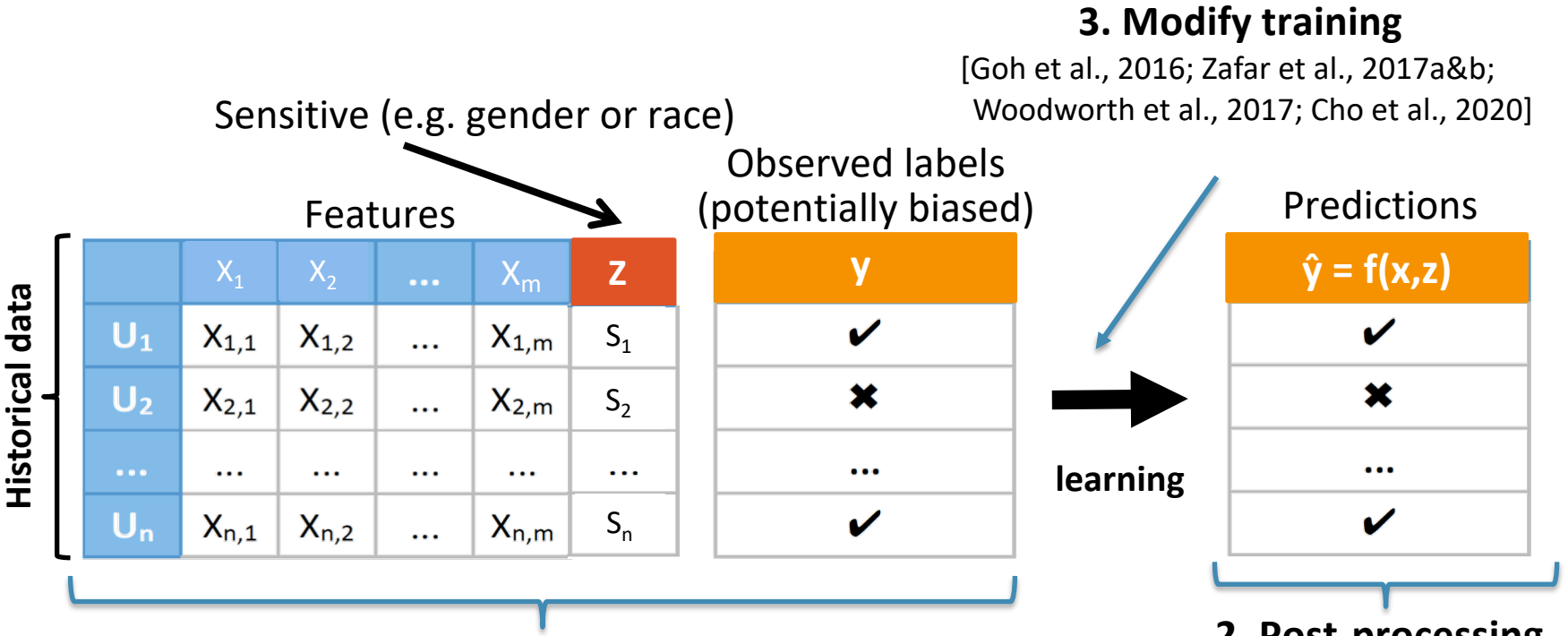
$$\hat{Y} \perp Z|Y$$

Sufficiency: Require the target and the sensitive features to be independent conditional on the prediction:

$$\hat{Y} \perp Z|\hat{Y}$$

[https://fairmlbook.org]

Fairness in ML - Mechanisms



Limitations & Open Questions

No universal definition

Different real-world scenarios...

Historical decisions
(potentially biased)

... require different definitions of fairness...

Parity (equality) in impact

... translates into into different measurements.

$$\hat{Y} \perp Z$$

No universal definition

Different real-world scenarios...

Ground Truth

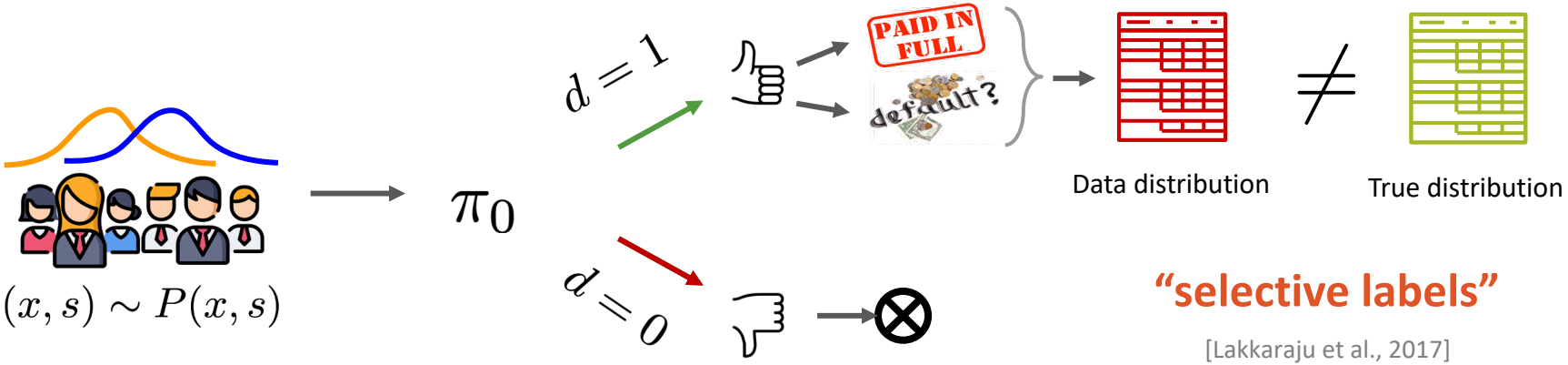
... require different definitions of fairness...

Parity (equality) in mistreatment or Equal Opportunity

... translates into into different measurements.

$$\hat{Y} \perp Z|Y$$

Wrong assumptions may amplify unfairness



"selective labels"

[Lakkaraju et al., 2017]

Wrong technical assumptions lead to "suboptimal predictive models that amplify unfairness beyond the levels exhibited by the (unfair) optimal policy" [Kilbertus et al., 2020]

When and how to use algorithms?

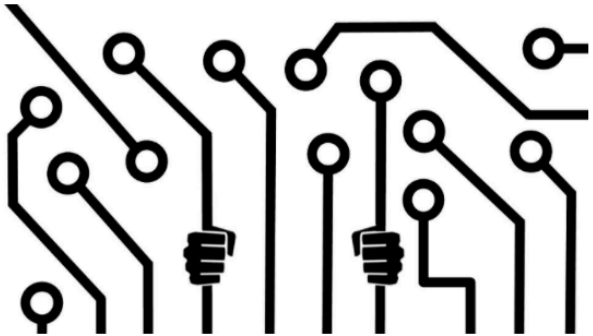
Essential to have a holistic view of the algorithm:

"Not because something is technically possible, it is the right thing to do."

Abolish the #TechToPrisonPipeline

Crime prediction technology reproduces injustices and causes real harm

Coalition for Critical Technology Follow
Jun 23 · 96 min read



Sign the letter—Email Springer—Authorship

Springer Publishing
Berlin, Germany
+49 (0) 6221 487 0
customerservice@springernature.com

RE: A Deep Neural Network Model to Predict Criminality Using Image Processing

June 22, 2020

When and how to use algorithms?

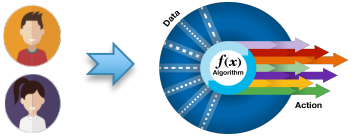
Essential to have a holistic view of the algorithm:

"Not because something is technically possible, it is the right thing to do."

- How to perform a fair data collection process?

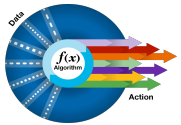
The New York Times

G.D.P.R., a New Privacy Law, Makes Europe World's Leading Tech Watchdog



- How are decisions made in practice?

OR/AND?



- Fairness beyond discrimination

Vox

Spend "frivolously" and be penalized under China's new social credit system

People who waste money on non-essentials or behave "badly" are penalized under the controversial new ranking.

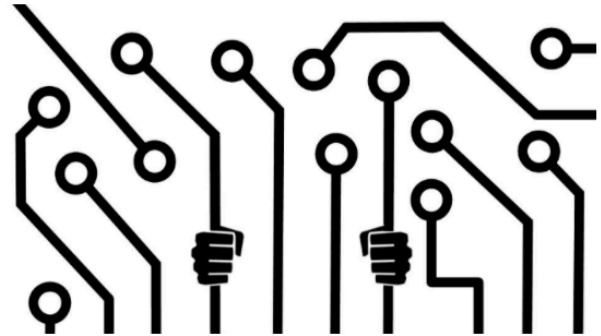
By Nadra Nittle | Nov 2, 2018, 6:50pm EDT

Abolish the #TechToPrisonPipeline

Crime prediction technology reproduces injustices and causes real harm

Coalition for Critical Technology Follow

Jun 23 · 96 min read

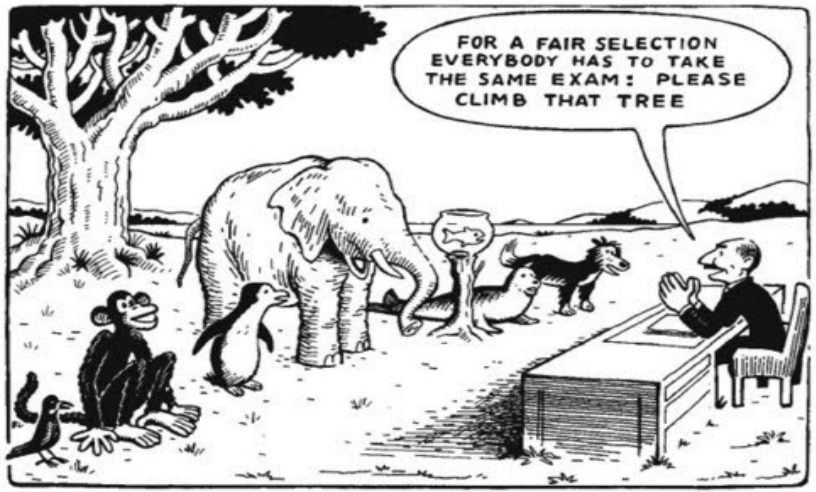


Sign the letter—Email Springer—Authorship

Springer Publishing
Berlin, Germany
+49 (0) 6221 487 0
customerservice@springernature.com

RE: A Deep Neural Network Model to Predict Criminality Using Image Processing

June 22, 2020



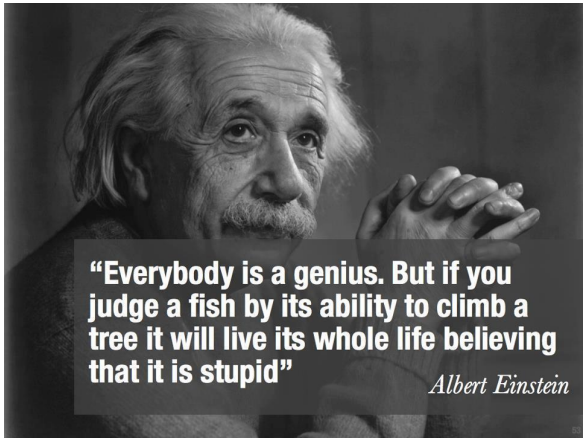
SIC Saarland Informatics Campus

Thanks!

Questions?

https://ivaleram.github.io/

ivalera@cs.uni-saarland.de



ellis
European Laboratory for Learning and Intelligent Systems