FORSCHUNG: VERÄNDERN

FORSCHUNGSTAG 2020

Fairness in Automated Decision-Making – FairADM

Frauke Kreuter, Ruben Bach, Christoph Kern

University of Mannheim

https://www.uni-mannheim.de/datascience/

Motivation

AI increasingly used in the public sector for high-stake decisions:

- Job training enrollment
- Social service intervention
- Detention and recidivism

AI systems may, however, **reinforce existing or creating new social inequalities and foster discrimination**

Algorithmic Fairness - Examples

ADM system used to predict recidivism risks in the U.S. criminal justice system systemically discriminates against black defendants

Amazon's hiring tool discriminated against women

➢ **Non-discrimination and fairness are key requirements for the trustworthy use of AI** in the European Union (High- Level Expert Group on Artificial Intelligence 2019)

Research Questions

Where is **ADM** used **in governmental contexts** in Germany?

How can these **ADM applications be classified**?

Which fairness notions should be considered when evaluating fairness?

Can we mitigate biases through technical solutions?

Approach

What is distributed? How many people can profit? Is it a **scarce resource**?

Is the **decision assistive or punitive**?

What is the **relative impact for the affected people**?

Use case: Evaluate fairness notions and constraints in a real-world application

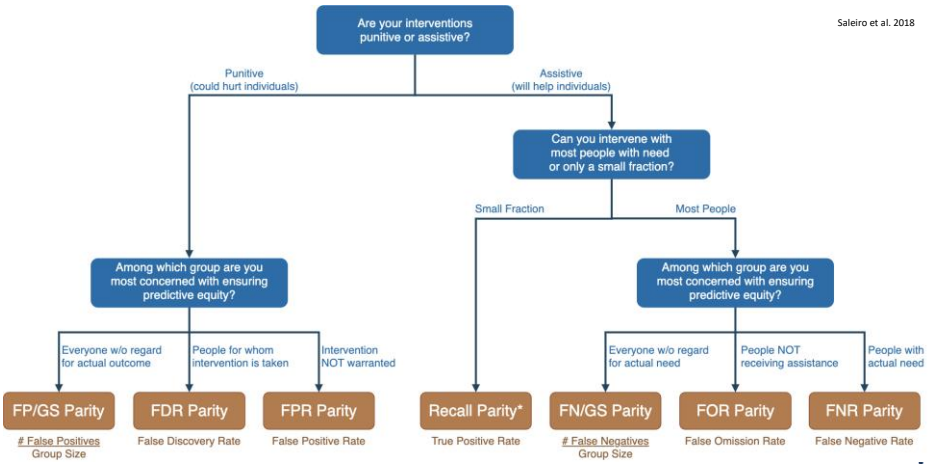
Administrative labor market records for Germany

Train re-employment prediction model

- **Fairness auditing** based on adequate fairness notions and sensitive attributes
- **Evaluate** and compare **bias correction methods**
- Investigate **long-term consequences**

Bias correction approaches (Berk et al. 2017)

- **Pre-processing**: Eliminating sources of unfairness in data before model training
- **In-processing**: Making fairness adjustments as part of the model building process
- **Post-processing**: Adjust model output post-training to make it more fair



Saleiro et al. 2018

Berk, R., Heidari, H., Jabbari, S., Kearns, M., Roth, A. (2017). Fairness in Criminal Justice Risk Assessments: The State of the Art. https://arxiv.org/abs/1703.09207

Saleiro, P., Kuester, B., Stevens, A., Anisfeld, A., Hinkson, L., London, J., Ghani, R. (2018). Aequitas: A Bias and Fairness Audit Toolkit. https://arxiv.org/abs/1811.05577

Baden-Württemberg Stiftung
WIR STIFTEN ZUKUNFT