Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Internationales Zentrum für Ethik in den Wissenschaften (IZEW)

Eberhard Karls Universität Tübingen

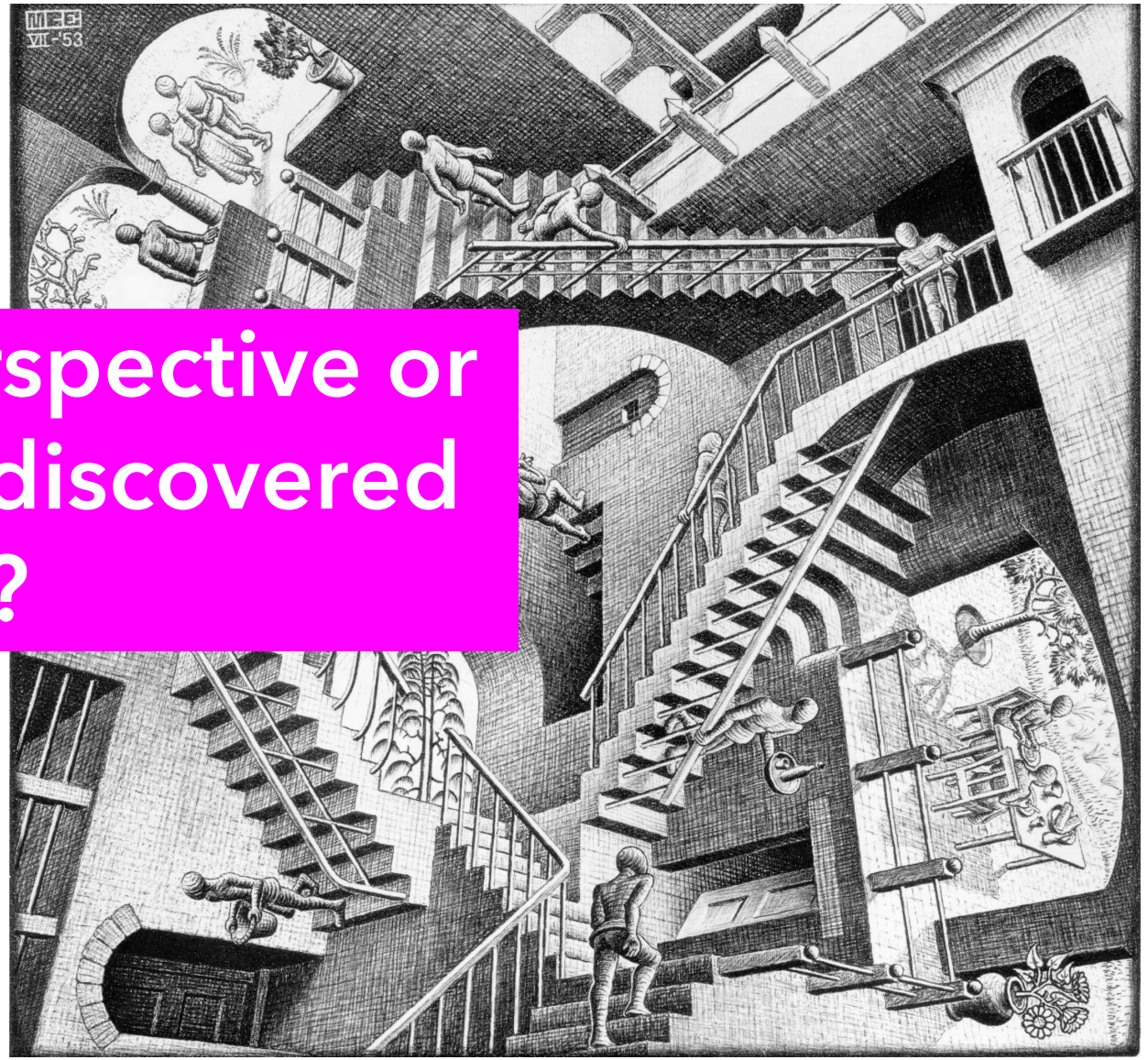
# Responsible Data Science

## Session 3: 27.04.2023, 15.15 – 18.30 h
## MA Seminar, SoSe 2023, Hasso-Plattner Institut

Dr. Simon David Hirsbrunner

# Today

15h15 Warm-up and recap
15h30 Student presentation COMPAS
16h00 Discussion
16h00 Introduction to key concepts in AI fairness
16h30 Discussion
16h50 --- Break ---
17h00 Introduction to Scenario-Based Design
17h15 Exercise in groups: building a value scenario
18h00 Sharing and discussing insights from groups
18h30 End




# A new perspective or issue you discovered yesterday?

# Recap

# Key concepts for today

- Bias
- Discrimination
- Discriminatory bias
- Fairness

# Bias in ML

Bias refers to a systematic error in a model's predictions, where the predicted values consistently differ from the true values due to flawed assumptions or data.

# Sources of bias in ML-based systems

*Examples*

- Representation Bias (missing subgroups in sample)
- Measurement Bias (mismeasured proxies)
- Presentation bias (user interface)
- Ranking bias (top-ranked results are clicked more often)
- Historical bias (existing imbalances affect user interaction)
- Behavioral bias (wrong translation across platforms)

(Mehrabi et al. 2021)

# Discrimination

Discrimination as "compounding historical injustice" (Helman 2018), according to which **disadvantage against members of socially significant groups** constitutes discrimination if it **reinforces historical injustices.**

(Behrendt and Loh 2022)


# Exercice: is this illegal discrimination in DE?

- Access to gym, restaurant, club denied with reference to disability.

Source: Antidiskriminierungsstelle


# Exercice: is this illegal discrimination in DE?

- Access to gym, restaurant, club denied with referen disability.


- Job offer not passed on with reference to too high age.

Source: Antidiskriminierungsstelle

# 1 Exercice: is this illegal discrimination in DE?

- Access to gym, restaurant, club denied with referen disability. illegal
- Job offer not passed on with reference to too high age illegal
- Requirement of a catholic religious affiliation for the job description for a director of a Catholic institution.

Source: Antidiskriminierungsstelle

# 1 Exercice: is this illegal discrimination in DE?

- Access to gym, restaurant, club denied with referen disability. — illegal
- Job offer not passed on with reference to too high age — illegal
- Requirement of a catholic religious affiliation for the job description for a director of a Catholic institution. — legal
- Hiring denied with reference to religious headscarf.

Source: Antidiskriminierungsstelle

# 1 Exercice: is this illegal discrimination in DE?

- Access to gym, restaurant, club denied with referen disability. — illegal
- Job offer not passed on with reference to too high age — illegal
- Requirement of a catholic religious affiliation for the job description for a director of a Catholic institution. — legal
- Hiring denied with reference to religious headscarf. — illegal
- A job as a model for youth fashion justifies the search for a person of a certain age.

Source: Antidiskriminierungsstelle

# 1 Exercice: is this illegal discrimination in DE?

- Access to gym, restaurant, club denied with referen disability. — illegal
- Job offer not passed on with reference to too high age — illegal
- Requirement of a catholic religious affiliation for the job description for a director of a Catholic institution. — legal
- Hiring denied with reference to religious headscarf. — illegal
- A job as a model for youth fashion justifies the search for a person of a certain age. — legal

Source: Antidiskriminierungsstelle

# Protected groups, examples

- Race
- Color
- Sex
- Language
- Religion
- Political or other opinion
- National or social origin
- Property
- Birth or other status

(according to Universal Declaration of Human Rights, Art. 2)

# Discriminatory bias in ML

Describes a situation where **AI systems exhibit unfair or unjustified discrimination** against certain groups or individuals based on their protected characteristics such as race, gender, age, religion, etc.

# Direct and indirect discrimination

- **Direct Discrimination.** Protected attributes of individuals explicitly result in non-favorable outcomes toward them.

- **Indirect Discrimination.** While individuals and groups appear to be treated based on seemingly unproblematic attributes, they still get to be treated unfairly as a result of implicit effects from their protected attributes (e.g. because of problematic proxies).

(Mehrabi et al. 2021: 10f)

# Example of indirect discrimination

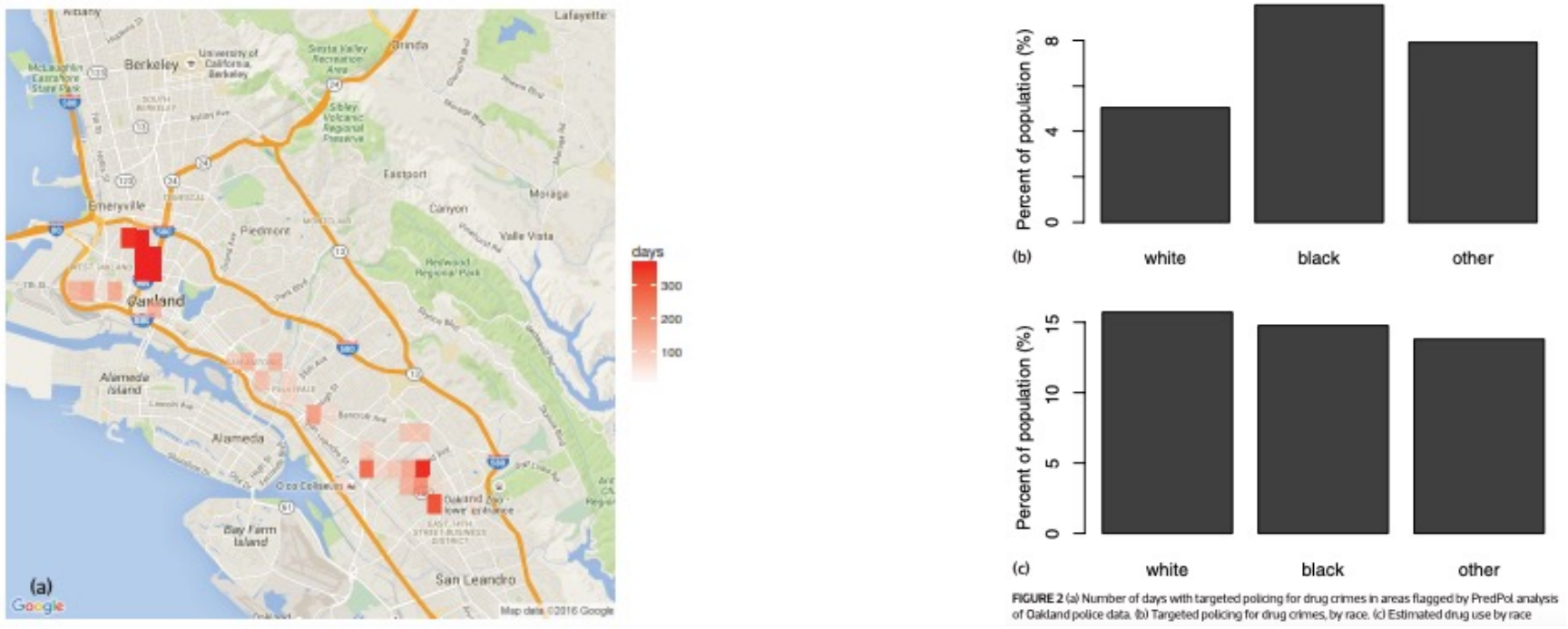



FIGURE 2 (a) Number of days with targeted policing for drug crimes in areas flagged by PredPol analysis of Oakland police data. (b) Targeted policing for drug crimes, by race. (c) Estimated drug use by race.

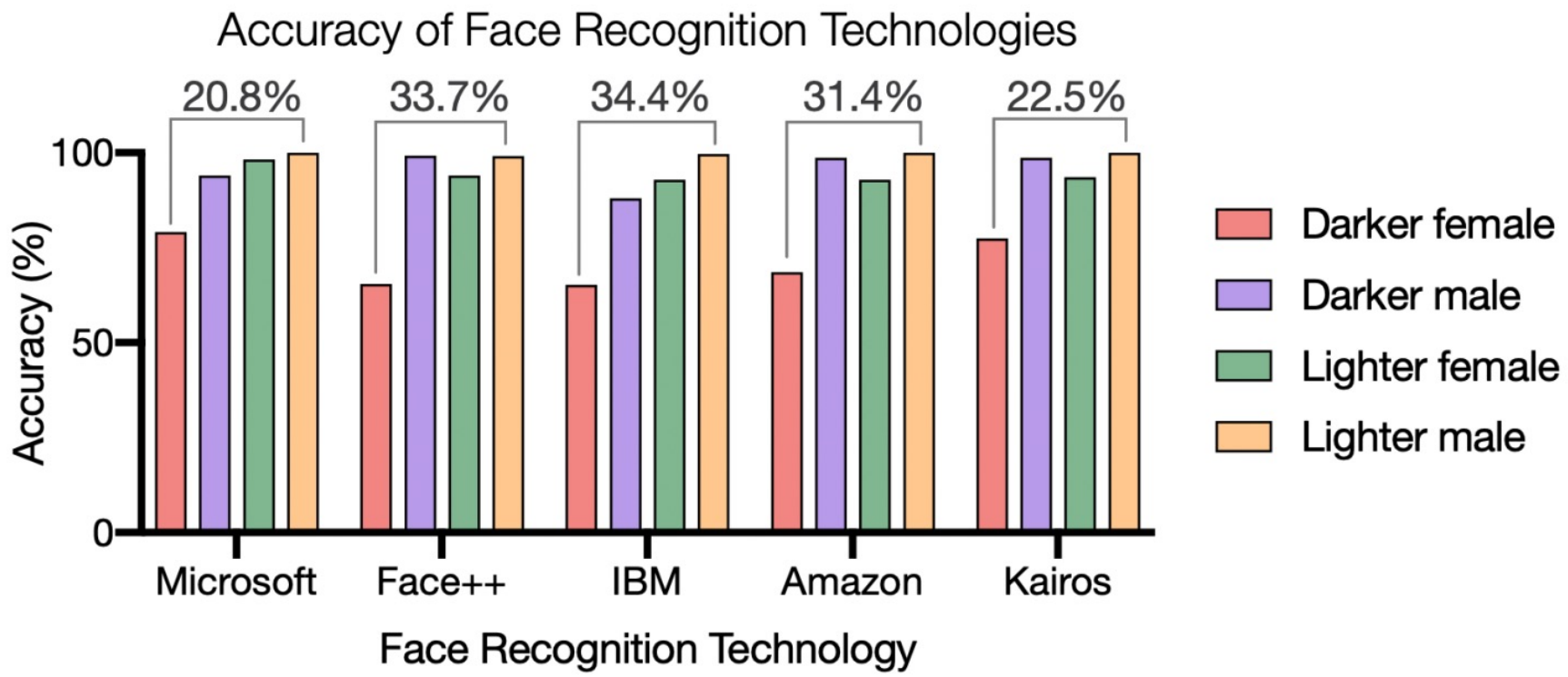
FIGURE 2 (a) Number of days with targeted policing for drug crimes in areas flagged by PredPol analysis of Oakland police data. (b) Targeted policing for drug crimes, by race. (c) Estimated drug use by race.

…

„[…] the more time police spend in a location, the more crime they will find in that location."

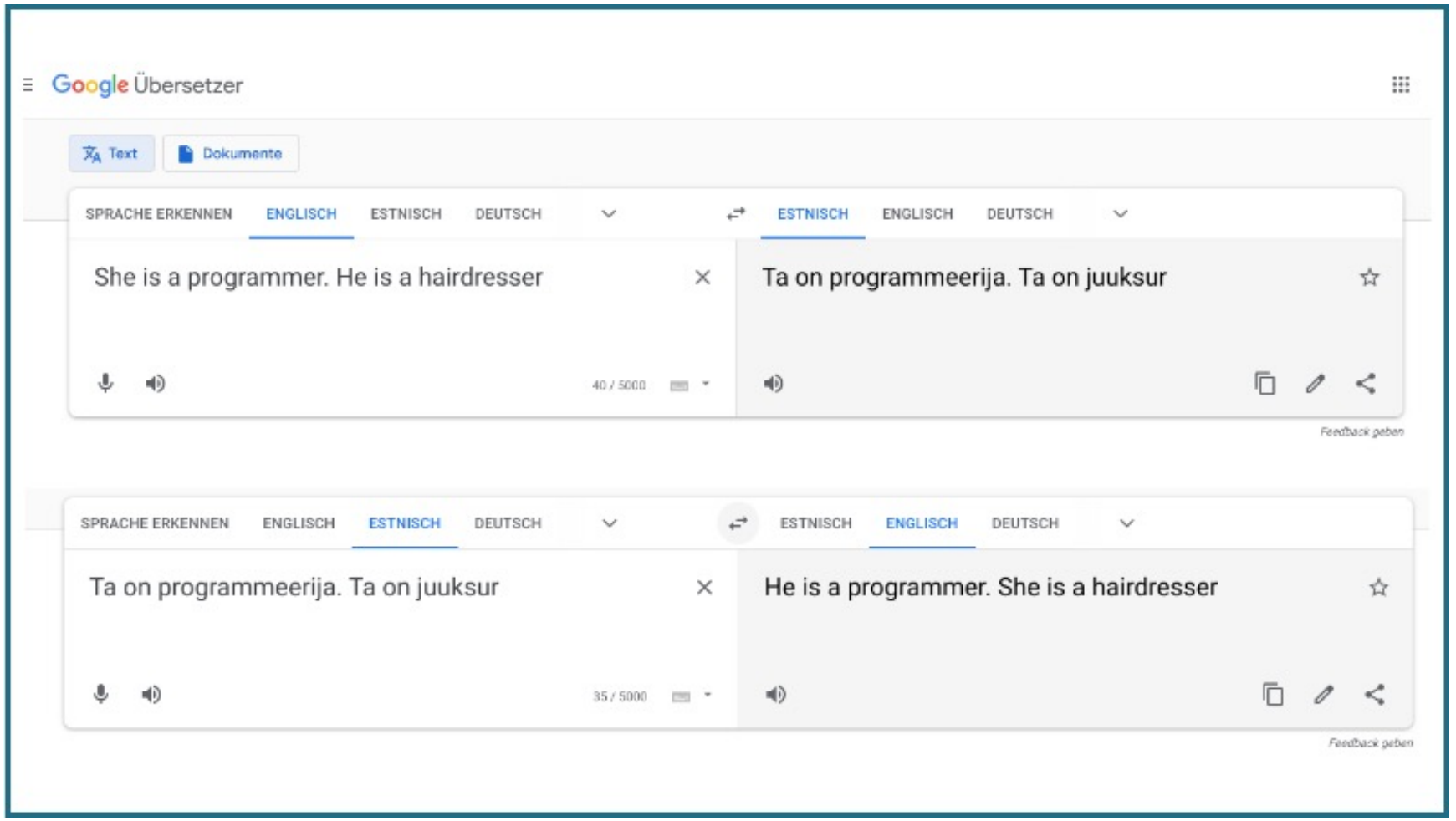
(Lum and Isaak 2016)

# Discriminatory bias in face recognition




https-//sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/.webp

# Discriminatory bias in NLP




https-//www.hiig.de/en/bias-in-natural-language-processing/.png; Bolukbasi et al. 2016)

# Fairness as an ethical value

Fairness describes a situation in which individuals and groups are treated equitably and impartially, without discrimination or favoritism based on arbitrary characteristics such as race, gender, religion, social status, or any other irrelevant factors. Fairness implies treating everyone with equal respect and dignity, and ensuring that decisions are based on objective criteria that are transparent and consistent.

- Fairness as ethical value and principle.
- Fairness is an instrumental value for the intrinsic value of justice.

# AI fairness

Freedom from bias in AI systems. AI fairness refers to the concept of **ensuring that AI systems and algorithms are designed and deployed in fair, unbiased, and non-discriminatory.** Technical approaches for AI fairness aim to prevent AI systems from perpetuating or amplifying biases that may exist in the data or the decision-making processes of the system.

# AI fairness: variants

**Group fairness** refers to the extent to which a machine learning model provides equitable outcomes for different groups of people (e.g. race, gender, age). A model that achieves group fairness aims to ensure that each group receives a fair and equitable outcome.

**Individual fairness** focuses on treating similar individuals similarly, regardless of their membership in any particular group. In other words, if two individuals have similar characteristics and behaviors, they should be treated similarly by the machine learning model.

**Subgroup fairness** aims to ensure that the model is fair not only for the overall population but also for subgroups within that population.

(Mehrabi et al. 2021: 13)

# Equalized odds

Equalized odds measures whether the algorithm is equally accurate across different groups, while controlling for the distribution of these groups in the population. Specifically, it measures whether the **true positive rate** and **false positive rate** for the algorithm are the same across all groups.

➢ **Focus on equal opportunity**

(Mehrabi et al. 2021: 11 f)

# Demographic parity

Demographic parity (or statistical parity) measures whether the algorithm is treating all groups equally in terms of the proportion of positive decisions (such as being hired or receiving a loan). The likelihood of a positive outcome should be the same regardless of whether the person is in the protected (e.g., female) group.

➢ **Focus on outcome distribution**

(Mehrabi et al. 2021: 12)

# Counterfactuals

The fairness technique counterfactuals examines hypothetical scenarios, or counterfactuals, in which different individuals or groups are treated differently and assessing whether the algorithm is making decisions without bias.

➢ **Focus on contextualization of biases**

(Kusner et al. 2017)

# Counterfactuals


(Cheong et al. 2023)

# Limits of algorithmic fairness

- Complete AI fairness (and freedom from bias) can only be approximated, not reached. It cannot become a feature of an AI system, but should rather be treated as a **desirable target.**

- To approximate algorithmic fairness, **multiple fairness perspectives/procedures** should be compared and/or linked.

- Algorithmic bias **may only arise during the operation** of the system. Therefore, fair AI cannot simply be established ex ante and has to be **reconsidered** upon changes in the data, system or social boundary conditions.

- AI fairness alone **doesn't automatically lead** to a dissolution of **systemic injustice** and an **advancement of diversity.** (cf. Verma and Rubin 2018; Barocas et al. 2021)

# Student presentation and discussion

Angwin J, Larson J, Mattu S, et al. (2016) **Machine bias:** There's software used across the country to predict future criminals. And it's biased against blacks. Propublica. Available at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

# Resources

**Gender shades project**
http://gendershades.org/overview.html

**Google developer website and resources**
https://developers.google.com/machine-learning/crash-course/fairness/video-lecture

**People + AI Research (PAIR)**
https://pair.withgoogle.com/explorables/measuring-fairness/

**Word bias** for NLP bias discovery
https://github.com/bhavyaghai/WordBias
http://130.245.128.219:6999/

# Sources

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. *Propublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Barocas, S., Hardt, M., & Narayanan, A. (2021). *Fairness and Machine Learning*. 253.

Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2477899

Behrendt, H., & Loh, W. (2022). Informed consent and algorithmic discrimination – is giving away your data the new vulnerable? *Review of Social Economy*, *80*(1), 58–84. https://doi.org/10.1080/00346764.2022.2027506

Bolukbasi et al. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In Proceedings of the International Conference on Advances in Neural Information Processing Systems.

Cheong, J., Kalkan, S., & Gunes, H. (2023). Counterfactual Fairness for Facial Expression Recognition. In L. Karlinsky, T. Michaeli, & K. Nishino (Eds.), *Computer Vision – ECCV 2022 Workshops* (Vol. 13805, pp. 245–261). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-25072-9_16

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual Fairness. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper/2017/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html

Lum, K., & Isaac, W. (2016). To Predict and Serve? *Significance*, *13*(5), 14–19. https://doi.org/10.1111/j.1740-9713.2016.00960.x

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, *54*(6), 1–35. https://doi.org/10.1145/3457607

Verma, S., & Rubin, J. (2018). Fairness definitions explained. *Proceedings of the International Workshop on Software Fairness*, 1–7. https://doi.org/10.1145/3194770.3194776

Images:
Slide 1: © Adobe Stock / kras99, Slide 3: Image by brgfx on Freepik