Responsible Data Science

Session 3: 15.05.2024, 13.30 – 16.45 h
MA Seminar, SoSe 2024, Hasso-Plattner Institut

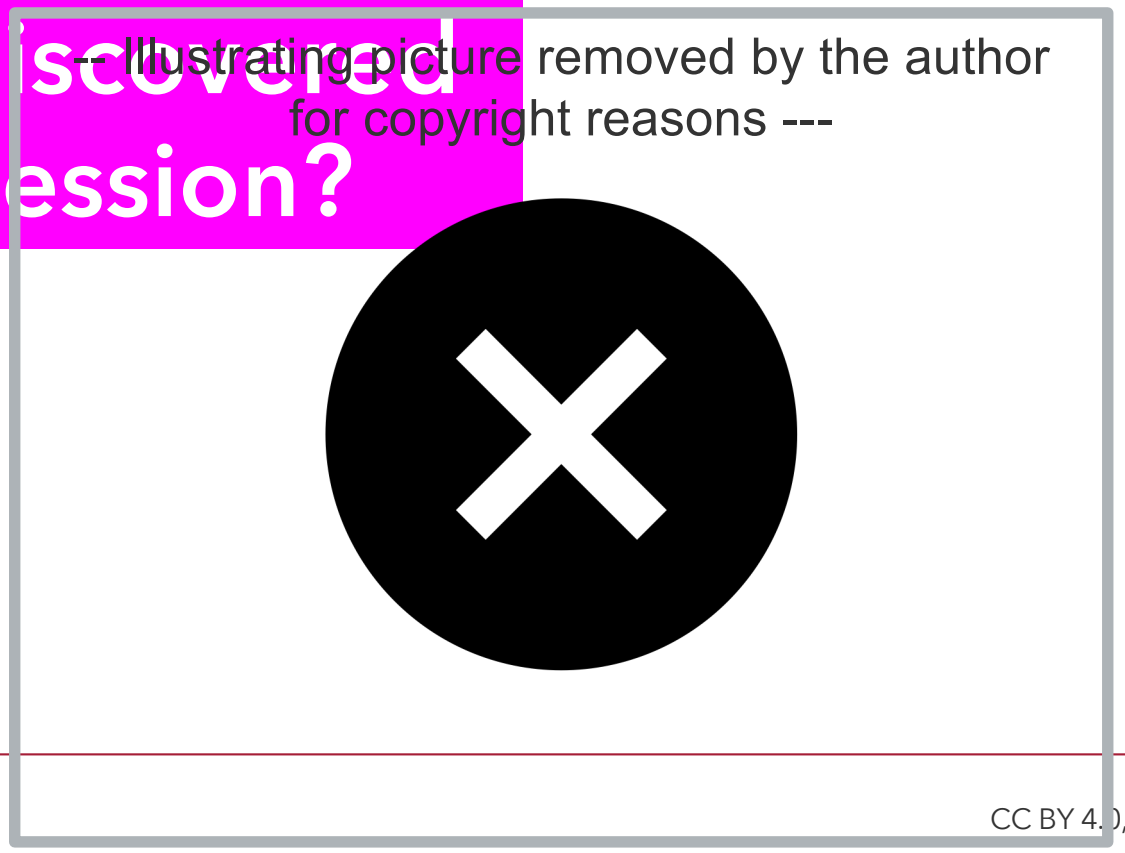
Dr. Simon David Hirsbrunner

Today

topic	time
Recap and warm-up	13h30
Student presentation COMPAS	14h00
Discussion	14h15
Introduction to key concepts in AI fairness	14h45
—— Break ——	15h15
Introduction to Scenario-Based Design	15h30
Exercise in groups: building a value scenario	15h50
Sharing and discussing insights from groups	16h20
Assignments next session	16h40
End	16h45

Recap

A new perspective or issue you discovered in the last session?



Ethical principles/codes in AI-related codices

Source: original research, based on Jobin et al. 2019; Hagendorff 2020

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.

Key concepts for today

- Discrimination
- Algorithmic discrimination
- Algorithmic fairness

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)

1 Exercise: is this legally considered discrimination in DE?

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

Source: Antidiskriminierungsstelle

1 Exercise: is this legally considered discrimination in DE?

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

illegal

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

Source: Antidiskriminierungsstelle

1 Exercise: is this legally considered 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 Exercise: is this legally considered 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 Exercise: is this legally considered 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 Exercise: is this legally considered 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

Bias refers to a systematic error in an AI model's predictions that leads to the discrimination against a group or individual based on their protected characteristics such as ra
ce, gender, age, religion, etc.

Note:
The kind of (discriminatory) bias discussed here is not to be confused with "inductive bias", which has a completely different meaning in ML.

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)

Discriminatory bias in ML

Describes a situation where **AI systems exhibit 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

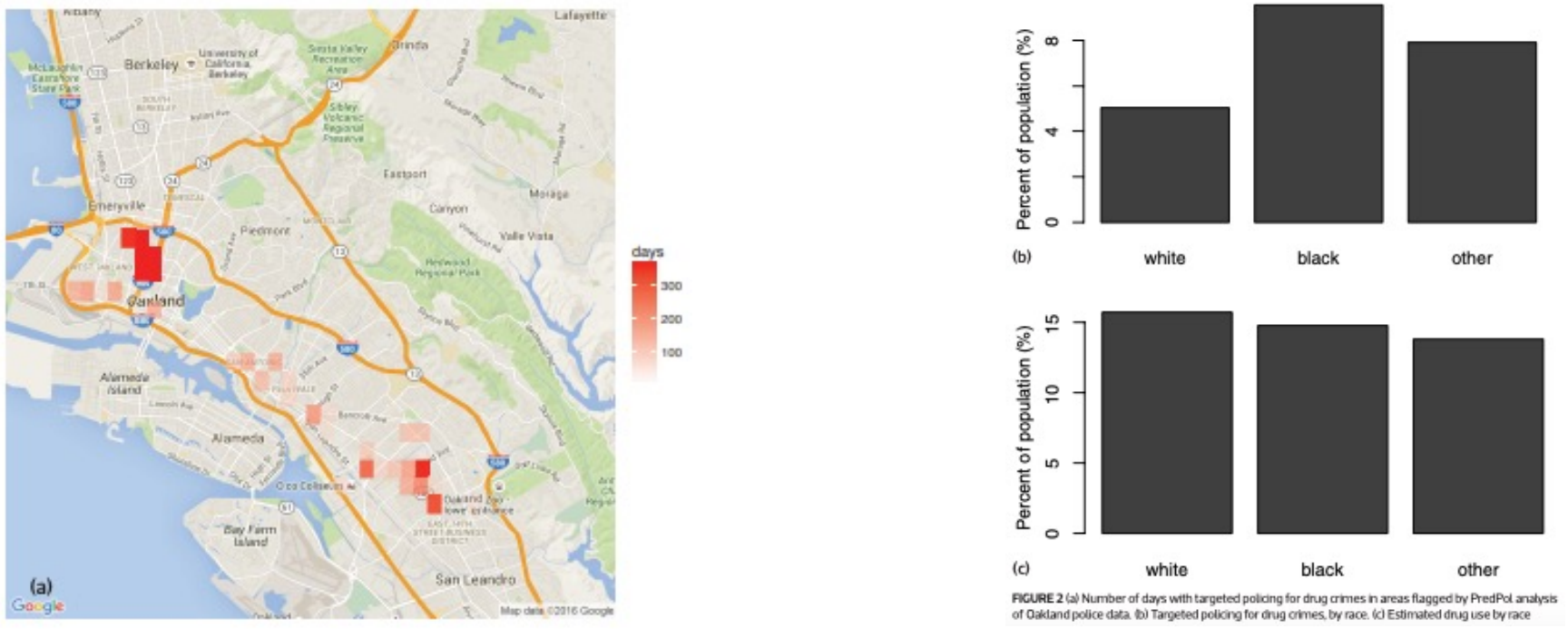


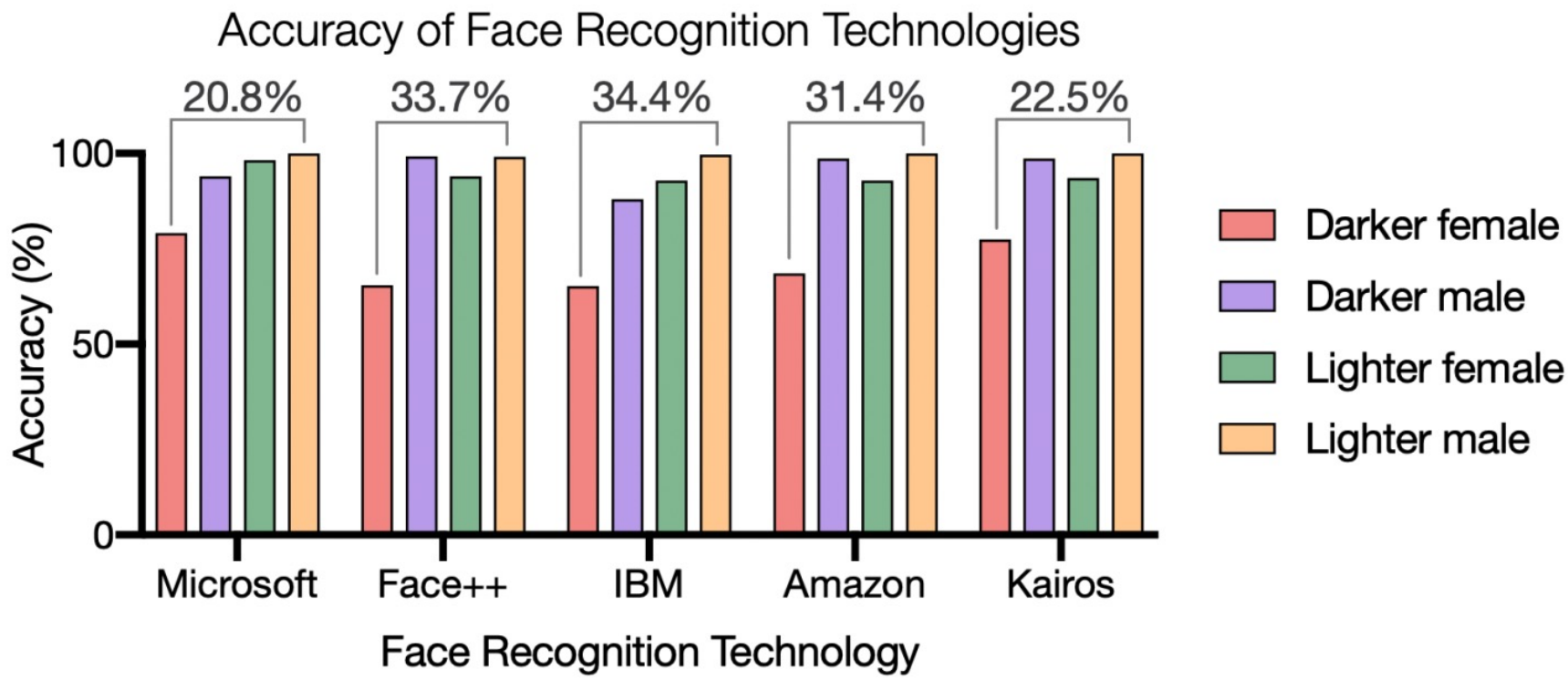
FIGURE 2 (a) Number of days with targeted policing for drug crimes in areas flagged by PredPol analysis of Oakland police data. (b) Targeted policing for drug crimes, by race. (c) Estimated drug use by race.

…

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

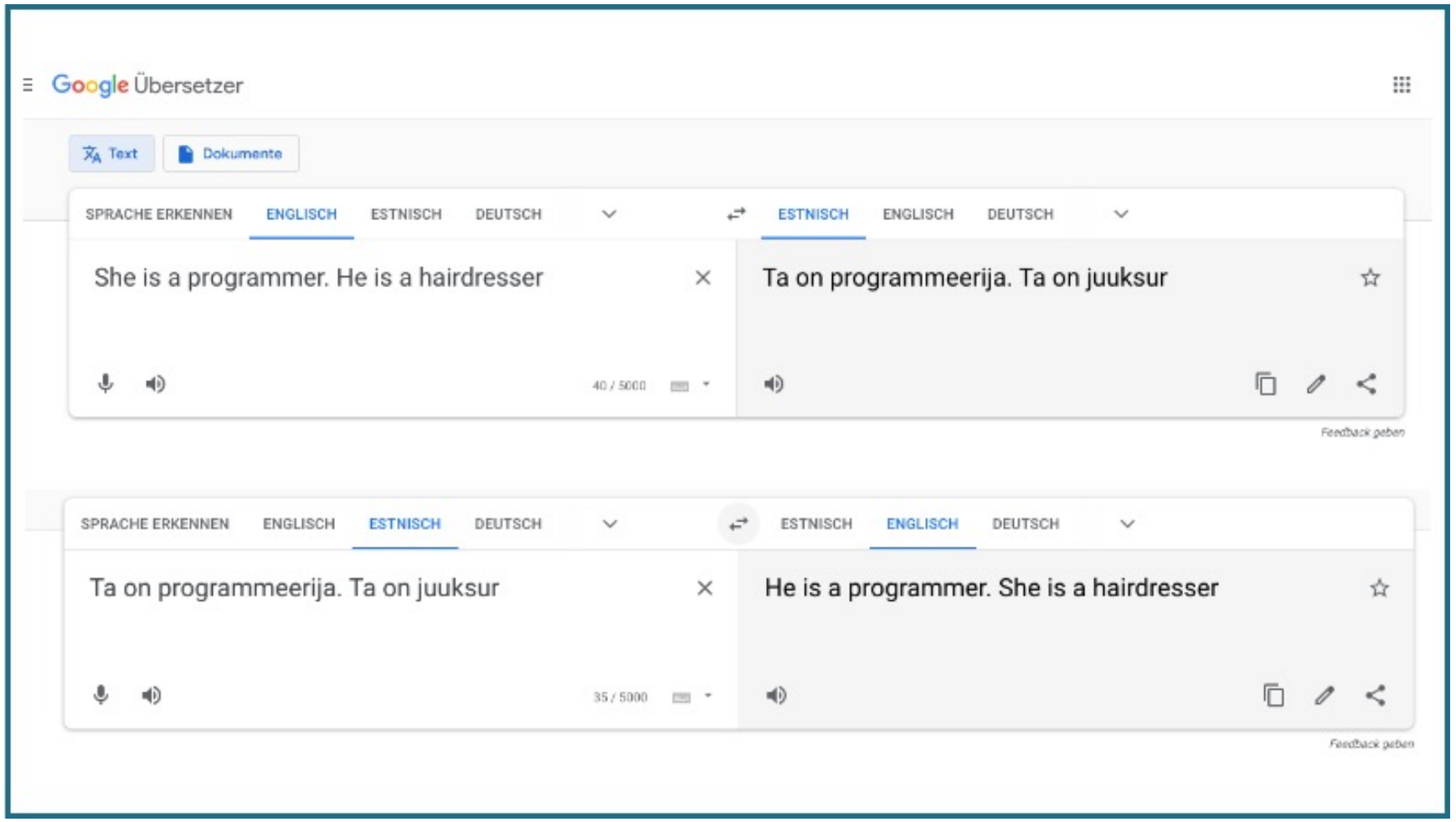
(Lum and Isaak 2016)

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 opportunities**

(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 equal outcomes (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)

AI fairness in LLMs

WIRED SECURITY POLITICS GEAR BACKCHANNEL BUSINESS SCIENCE CULTURE IDEAS MERCH

TOM SIMONITE BUSINESS DEC 8, 2020 4:39 PM

Behind the Paper That Led to a Google Researcher's Firing

Timnit Gebru was one of seven authors on a study that examined prior research on training artificial intelligence models to understand language.

Timnit Gebru, a prominent artificial intelligence researcher, says she was fired after refusing to retract or take her name off an academic paper. PHOTOGRAPH: CODY O'LOUGHLIN/REDUX

Bender EM, Gebru T, McMillan-Major A, et al. (2021) On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency,

https://www.wired.com/story/behind-paper-led-google-researchers-firing/

Limits of algorithmic fairness

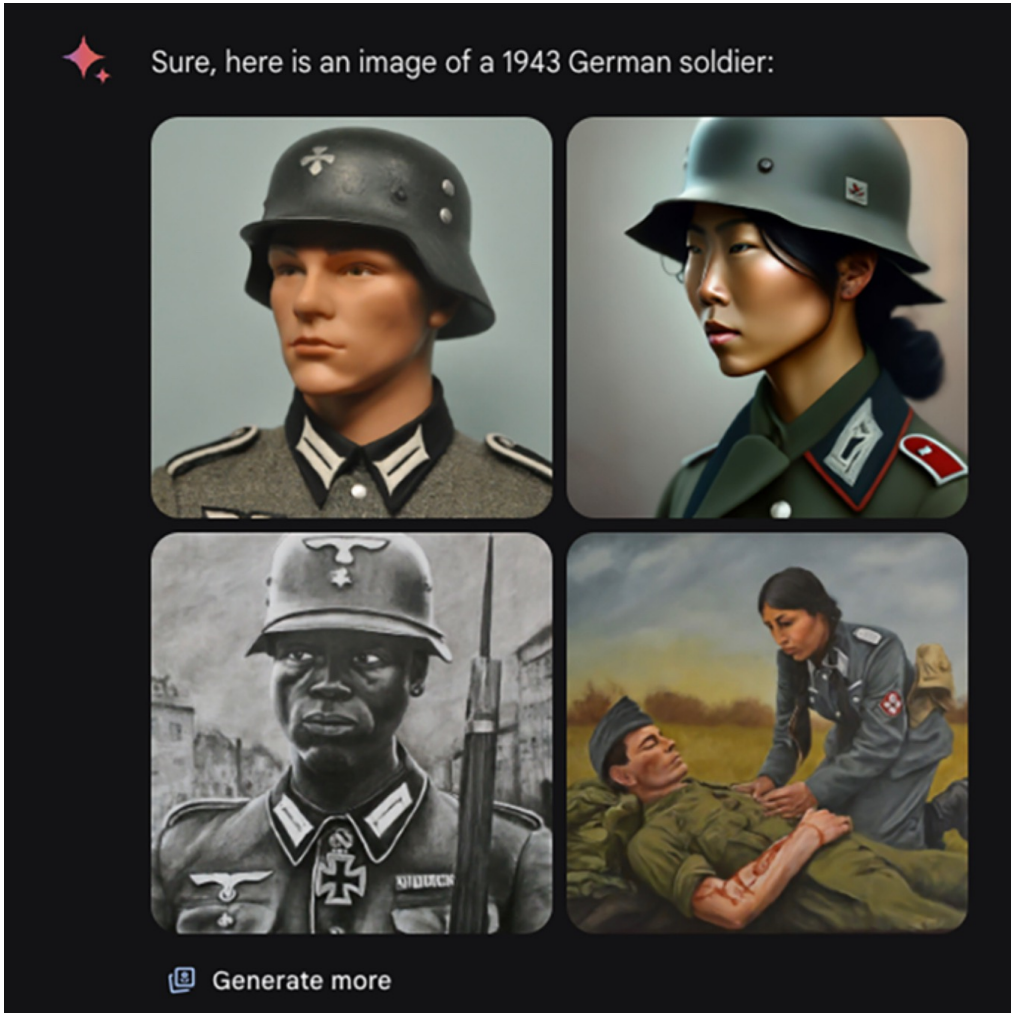
- Complete AI fairness (and freedom from bias) can only be approximated, not reached. It cannot become a feature of an AI system, but should rather be treated as a **desirable target.**

- 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)

Questions of diversity



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

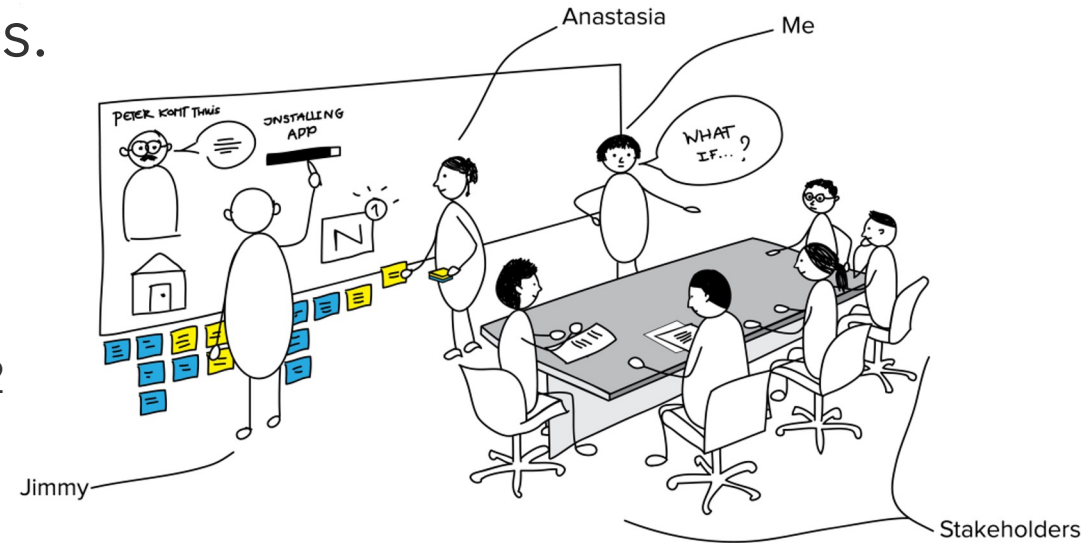
People + AI Research (PAIR)
https://pair.withgoogle.com/explorables/measuring-fairness/

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

Scenario-based design (I): definition

Scenario-based design is a family of techniques in which the **use of a future system is concretely described at an early point in the development process.** Narrative descriptions of envisioned usage episodes are then employed in a variety of ways to guide the development of the system that will enable these use experiences.

(Rosson and Carroll 2007: 1)
image: https://unitid.nl/2014/02/scenario-based-design/



Scenario-Based Design (II): ingredients

- **Actors** (direct and indirect stakeholders)
- **Setting** (application context, situation of use)
- **Tools and objects** (technologies, interfaces)

(Rosson and Carroll 2007)

Scenario-Based Design (III): example

S Martin is a police officer at the organized crime unit of the federal police. He currently investigates the selling of fake COVID-19 vaccination passports by an alleged criminal organization named *The Medics*. The Medics offer the counterfeit certificates to their S customers via the Telegram messenger. Unknown to Martin yet, S Chris, S Carlos, and S Eggert are Medics members, also communicating with their colleagues and suppliers via group Telegram channels while using pseudonyms, sometimes coded language, and images. In their free time, they also communicate with several friends, including their girlfriends, S Sarah and S Marta, who are unaware of their business. Martin's police unit gathers much information about The Medics using traditional investigative methods. This information leads to the identification of the suspect, Chris, who seems to be a low-level member of The Medics. On one evening, Chris is found with blank vaccination certificates during a traffic stop. He is arrested, and his phone is seized by investigator Martin, who aims at using the information on the phone to track down the individuals pulling the strings. After calling judge S Robert to get a search warrant, which is granted, he then searches Chris's unprotected phone, finds the Telegram communication, and extracts it. He recalls that his superior, S Dr. D, asked him to try out the new AutoCommAnalyzer software, which was recently purchased from the multinational company AI-Tech Corp. The software purchase was part of a strong push by the government to digitally optimize work processes at the police forces. Martin looks at the training notes by the head developer S Molly, trying to remember how the machine learning-driven software — trained with texts by S Alf and S Bert — is supposed to direct him to the relevant communication. The software presents him with the most frequent contacts, with Sarah on top. He reads through this communication, as the software has flagged several words like package and hospital, discovering some explicit images but finding that the

(Fischer MT, Hirsbrunner SD, Jentner W, et al. (2022) Promoting Ethical Awareness in Communication Analysis: Investigating Potentials and Limits of Visual Analytics for Intelligence Applications. In: *2022 ACM Conference on Fairness, Accountability, and Transparency*, New York, NY, USA, 20 June 2022, pp. 877–889. FAccT '22. Association for Computing Machinery. DOI: 10.1145/3531146.3533151)

Scenario-Based Design (IV): qualities

- Scenarios are concrete but rough;
- Scenarios maintain an orientation to people and their needs;
- Scenarios are evocative, raising questions at many levels.

(Rosson and Carroll 2007)

Value scenario

- Stakeholders
- Pervasiveness
- Time
- Systemic effects
- **Value implications**

Value scenario = VSD + Scenario-Based Design (SBD)

(Nathan et al. 2007)

2 Exercise: creating a value scenario

- Split into small groups
- Pick one of the following use cases:
  - AI-supported facial recognition for migration control
  - AI-supported dating platform
  - AI-supported recruitment tool
- Imagine an AI system designed for fairness (visual brainstorming)
- Report some key insights to the big group

See you on 22.05.!

simon.hirsbrunner@uni-tuebingen.de

-- Illustrating picture removed by the author for copyright reasons ---

Readings:

Taylor, L., Floridi, L., & van der Sloot, B. (2017). Introduction: A New Perspective on Privacy. In L. Taylor, L. Floridi, & B. van der Sloot (Eds.), Group Privacy: New Challenges of Data Technologies (pp. 1–12). Springer International Publishing. https://www.stiftung-nv.de/sites/default/files/group-privacy-2017-authors-draft-manuscript.pdf

Mühlhoff R (2023) Predictive privacy: Collective data protection in the context of artificial intelligence and big data. Big Data & Society 10(1). DOI: 10.1177/20539517231166886. https://journals.sagepub.com/doi/epub/10.1177/20539517231166886

Bibliography

The entire bibilography for the course can be found on Github here: https://github.com/simonsimson/responsible-data-science/blob/main/slides/Bibliography-of-the-entire-course.pdf

Image sources

Most sources are cited on the relevant slide. Slide 1: © Adobe Stock / kras99