Bias in Daten und Algorithmen in Text und Arbeit: alte Probleme, neue Ansätze

Bettina Berendt
TU Berlin, Weizenbaum-Institut, und KU Leuven
www.berendt.de/bettina
ZiF-Konferenz 2022, Bielefeld
21.10.2022

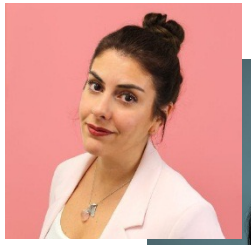
Thank you, team!

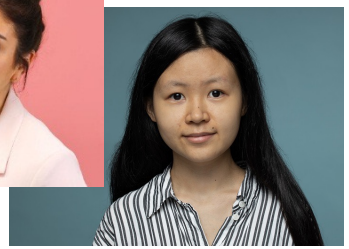
Sonja Wang, Mila Miceli, Tianling Yang, Kristen Scott, Pieter Delobelle

- Chair for Internet and Society @ Faculty of Electrical Engineering and Computer Science



- (TU) Director of the Weizenbaum Institute for the Connected Society
- Interdisciplinary Research on Digitalization and Society
- Policy Advice





- Guest professor in the DTAI group @ Department of Computer Science
- Research projects:
  - NoBIAS



  - VeriLearn
  - FLAIR (Flemish AI Programme)

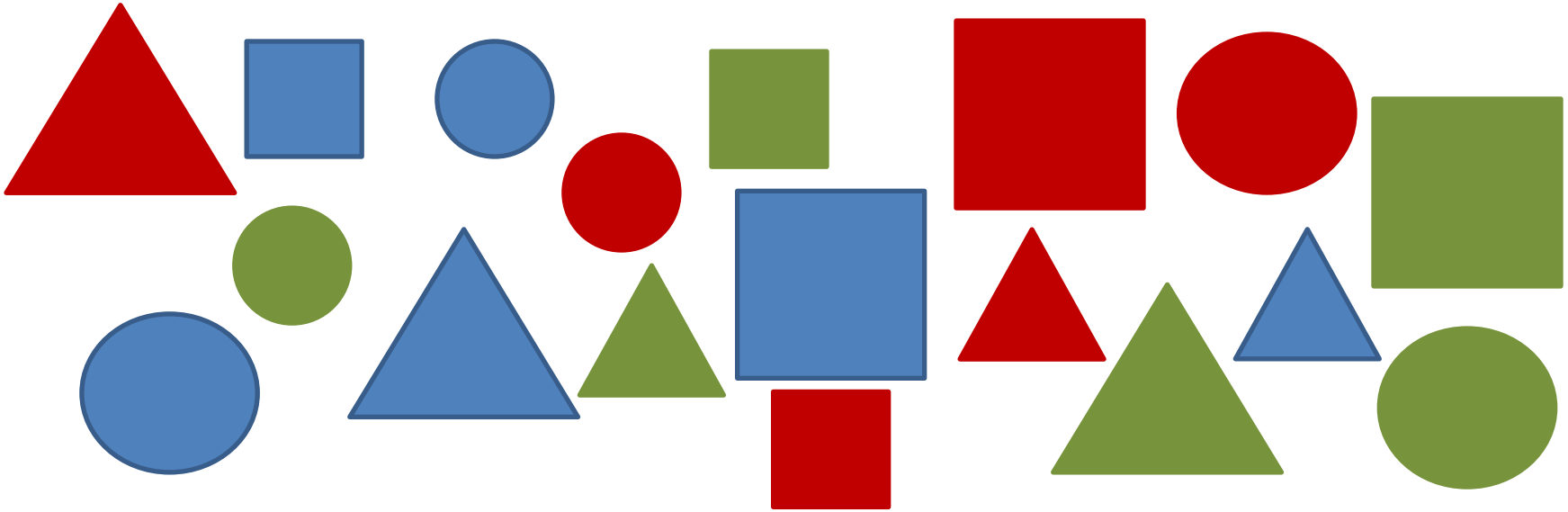


DIAMOND

Bias und Diskriminierung

Menschen

- sind und verhalten sich unterschiedlich (?)
- werden unterschiedlich behandelt



Wann ist das Diskriminierung?

Diskriminierung und Bias: Rechtliches und Sozialpsychologisches

- Benachteiligungen aus Gründen der Rasse oder wegen der ethnischen Herkunft, des Geschlechts, der Religion oder Weltanschauung, einer Behinderung, des Alters oder der sexuellen Identität (§1 AGG)
- Niemand darf wegen seines Geschlechtes, seiner Abstammung, seiner Rasse, seiner Sprache, seiner Heimat und Herkunft, seines Glaubens, seiner religiösen oder politischen Anschauungen benachteiligt oder bevorzugt werden. Niemand darf wegen seiner Behinderung benachteiligt werden. (Art. 3 GG; s.a. Art. 21 EU-Charta)
- mittelbare Benachteiligung [:] dem Anschein nach neutrale Vorschriften, Kriterien oder Verfahren Personen [...] benachteiligen können, es sei denn, [...] rechtmäßiges Ziel sachlich gerechtfertigt und die Mittel [...] angemessen und erforderlich. (§3 (2) AGG)

- Bias *(neudeutsch):* Voreingenommenheit, Vorurteil

Diskriminierung und Bias: In algorithmischen Systemen und Statistik

- “A system discriminates unfairly if it denies an opportunity or a good or if it assigns an undesirable outcome to an individual or group of individuals on grounds that are unreasonable or inappropriate.”
- A computer system has bias if it “systematically and unfairly discriminates against certain individuals or groups of individuals in favor of others”.

(Friedman and Nissenbaum, 1996)

- Bias in der Statistik und in naturwissenschaftlichen Kontexten: systematischer Fehler (viele verschiedene Typen)

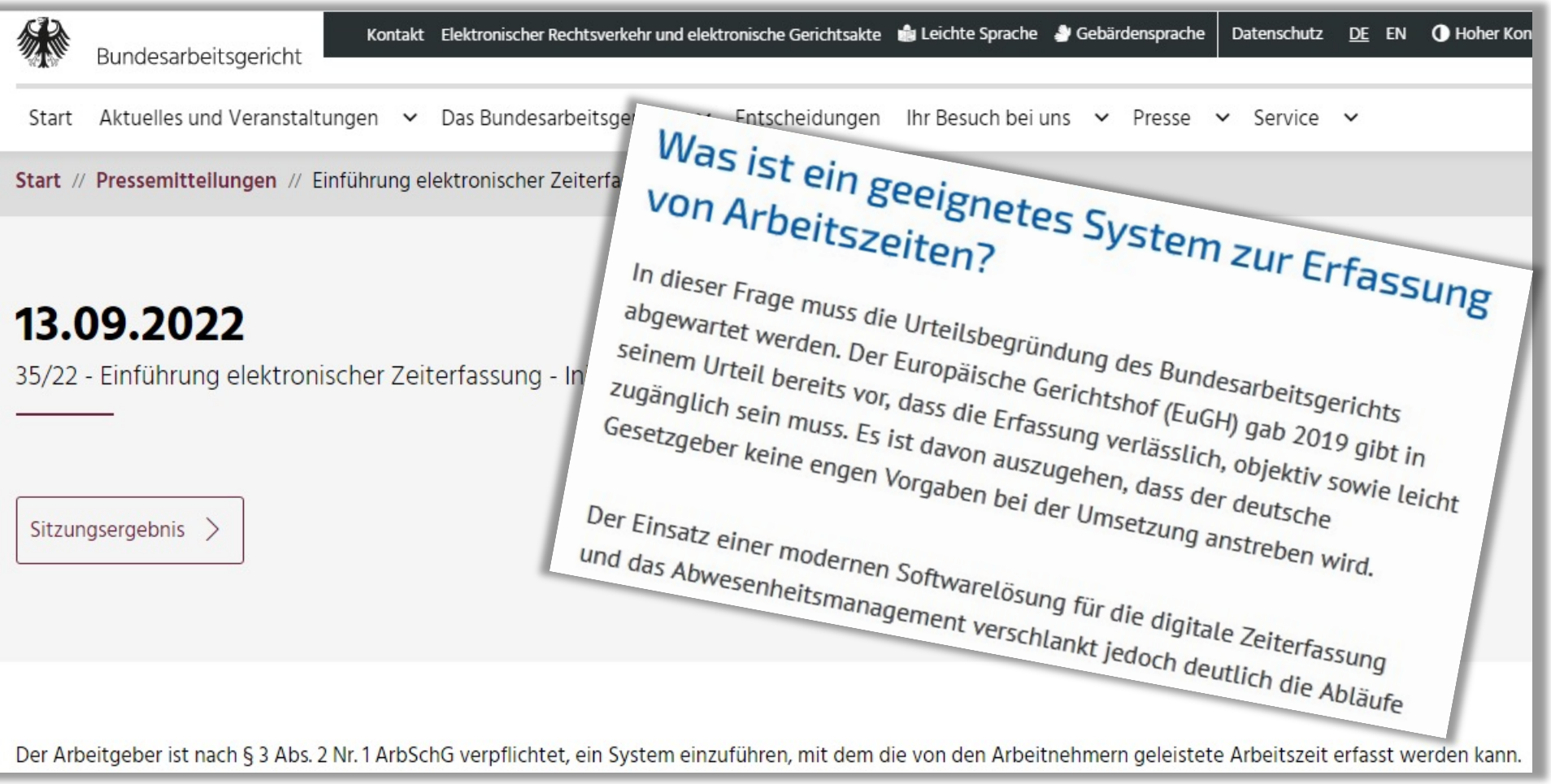
Was fordert und was behauptet die KI-Verordnungs-”Gesetzesfamilie”?

- DSGVO – Erwägungsgründe
  - ~ Persönliche Daten sollten **so** verarbeitet werden, **dass** keine diskriminierenden Effekte entstehen. (71, 75, 85)
- KI-Verordnung – Erwägungsgründe
  - ~ **Weil** bestimmte KI-Systeme diskriminierende Effekte haben können, sollten sie nur unter Auflagen betrieben werden können – oder ganz verboten werden. (15, 17, 28, 33, 35-39, 44, 47)

- Leider fehlt diesen Gesetzen ein Verständnis dessen, wie algorithmische Diskriminierung funktioniert und Prinzipien, die ihr entgegenwirken (=/= Datenschutz für Privatsphäre)
- Ziel dieses Vortrags: ein solches Verständnis in 30 Minuten ☺

Berendt (in Rostalski & BMU 2022)

Beispiel



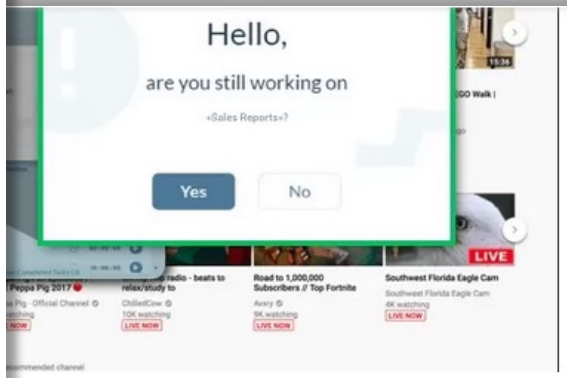
https://www.bundesarbeitsgericht.de/presse/einfuehrung-elektronischer-zeiterfassung-initiativrecht-des-betriebsrats/
https://zmi.de/arbeitszeiterfassung-pflicht-eugh-urteil/

Liegt ja nahe, oder …?

Top 10 Software To See If Remote Employees Are Working



Worried about whether your remote workforce is working or not?



Time Use report that details what sites an employee accessed and

employees' screen at random intervals to ensure that they're on

an be disabled at any time.

delete a screenshot in case it contains any sensitive information.

C) Inactivity Tracker

Time Doctor ensures that your staff is actively working during their work hours through the in-built inactivity tracker.

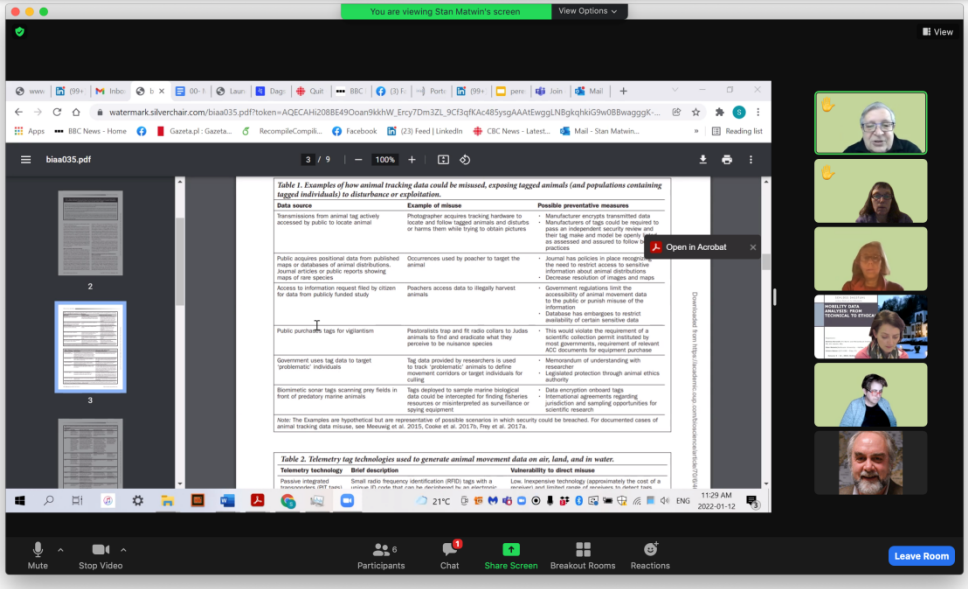
When there's been no keyboard or mouse activity for three minutes, Time Doctor automatically assumes that the worker is on a break and pauses the timer.

https://biz30.timedoctor.com/remote-employee-software/

Gedankenexperiment *(nur leicht fiktiv *)*

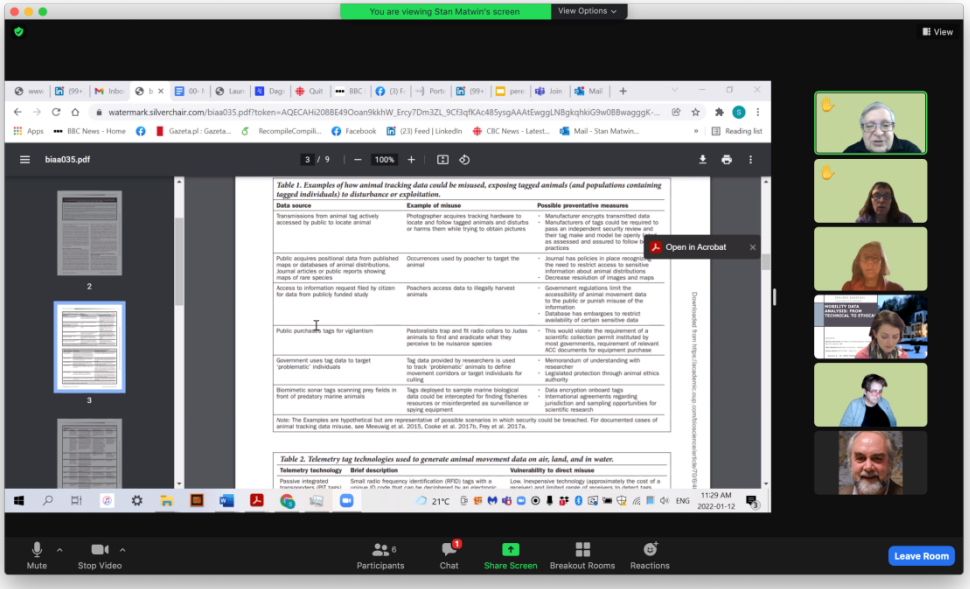
For example, earlier this year, Pricewaterhouse Coopers ("**PwC**") developed an FRT-based tool that would allow clients to track their employees whilst they worked from home. The system reportedly uses employees' webcams to log absences from their desks and forces them to give a written explanation for time spent away from their computer screens. PwC said in a statement that they are "*developing [the] technology specifically to support the compliance environment required for traders and front office staff in financial institutions. Crucially it is designed to support those adhering to the regulations while remote working, in the least intrusive, pragmatic way.*"

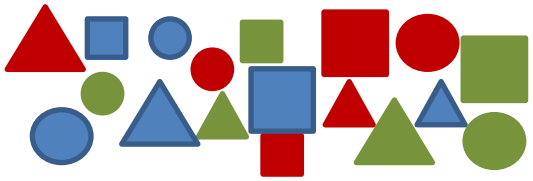
More recently we have seen companies considering pairing thermal cameras with facial recognition software and personnel directories as part of their COVID-19 health & safety protocols



*) weniger fiktiv z.B. für Arbeiter bei Amazon (Williams, Miceli, & Gebru, Noema Magazine 2022)

Es könnte passieren, dass …





… ein KI-Zeiterfassungssystem die Aktivität *blauer Personen* sehr fehlerhaft registriert. *

Was würden Sie einfordern?

Was würden Sie wissen wollen?

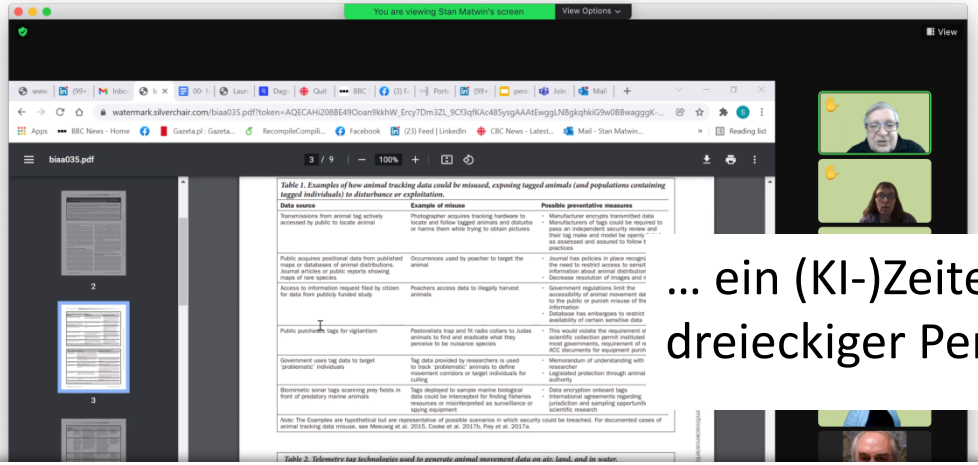
*) Inspiriert von (Buolamwini & Gebru, FAT 2018)

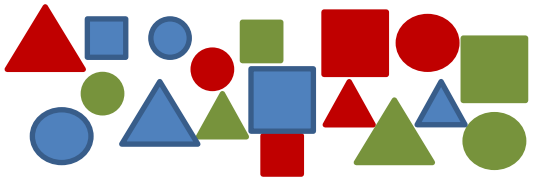
→ Daten

➔ Bias (und mögliche Diskriminierung) durch

- Daten
  - Selektionsbias (wurde hier beleuchtet, es gibt natürlich weitere Risiken)
  - Ein Ansatz: repräsentativer sampeln

Was würden Sie einfordern, wenn …





… ein (KI-)Zeiterfassungssystem die *Tätigkeiten* dreieckiger Personen sehr fehlerhaft registriert?

Time Doctor ensures that your staff is actively working during their work hours through the in-built inactivity tracker.

When there's been no keyboard or mouse activity for three minutes, Time Doctor automatically assumes that the worker is on a break and pauses the timer.

This

Worried that your employees maybe on social media and instant messaging sites when they're supposed to be working?

With Time Doctor, you don't have to worry about that.

Whenever an employee accesses unproductive sites like these, the app automatically sends them a pop-up asking them if they're still working. This little nudge is usually enough to get them off the social media site and back to work.

Ein nur wenig fiktiver Vorgriff: KI-Bildinterpretation statt Webseiten-Blacklists und Keyboard-Tracking

- KI-Bildinterpretation braucht (von Menschen) gelabelte Beispielbilder, hier von “produktiver Tätigkeit” und “unproduktiver Tätigkeit”
- Solche Annotationsaufgaben finden i.d.R. in Niedrig(st)-Lohnarbeit und unter starren Vorgaben, wie zu labeln ist, statt.
- Welches/wessen Verständnis von “produktiver Tätigkeit” würde dann in den Daten repräsentiert sein?

Miceli et al. (ACM HCI 2020), Miceli, Posada, & Yang (ACM HCI 2022)

→ Repräsentation

➔ Bias (und mögliche Diskriminierung) durch

- Daten
  - Selektionsbias (wurde hier beleuchtet, es gibt natürlich weitere Risiken)
  - Ein Ansatz: repräsentativer sampeln
- Repräsentation
  - Reporting bias and social desirability bias; Normalisierung gesellschaftlicher Machtverhältnisse
  - Ein Ansatz: Stakeholder-Einbindung / Co-Design in Prozessen der Wissensgenerierung

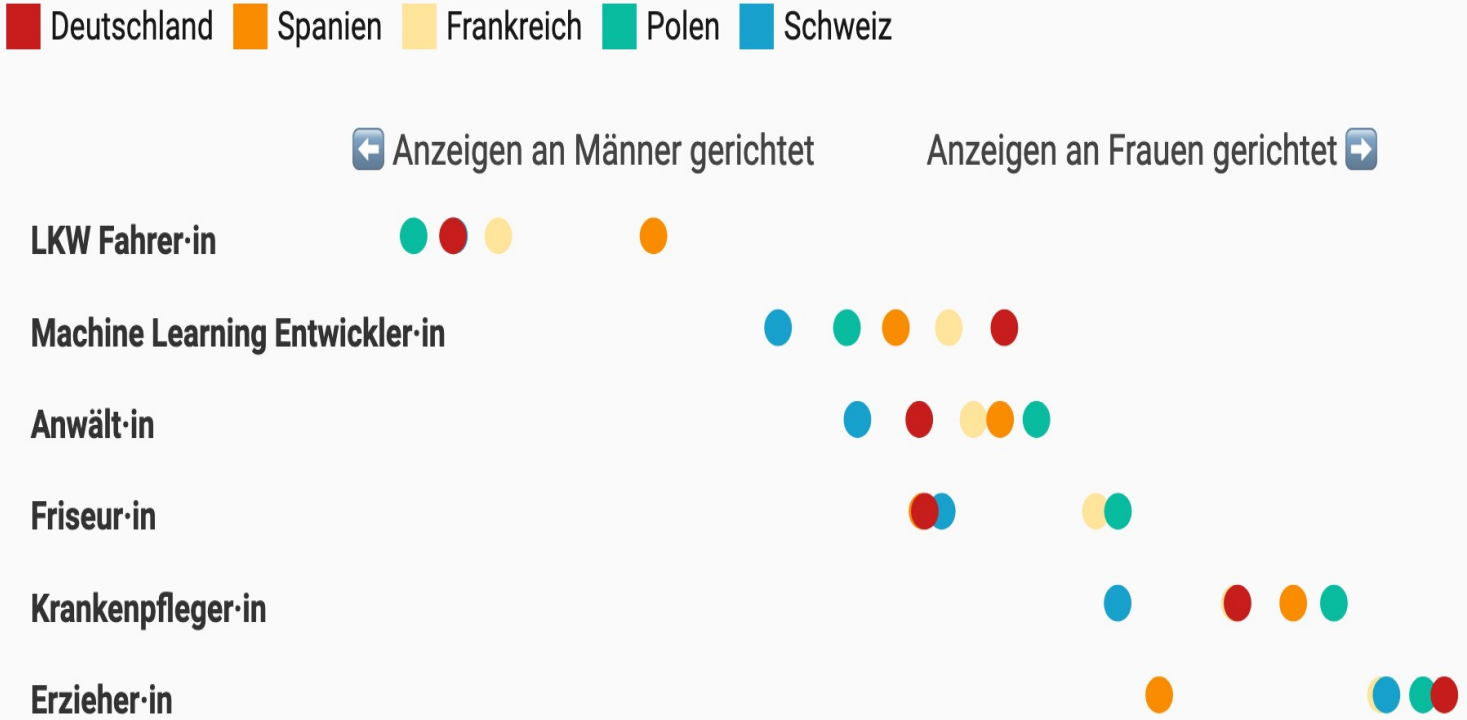
Scott, Wang, Miceli, Delobelle, Sztandar-Sztanderska, & Berendt (FAccT 2022)

Entscheidungen

Facebook verwendet grobe Stereotypen, um die Anzeigenschaltung zu optimieren

Algorithmic Decision Making

Wir haben Anzeigen für sechs verschiedene Stellenangebote in fünf Ländern gekauft. So optimierte Facebook die Anzeigenimpressionen, aufgeschlüsselt nach Geschlecht.



Die Grafik basiert auf 102.472 Anzeigenimpressionen zwischen 27.8. und 3.9.

Chart: AlgorithmWatch • Get the data • Created with Datawrapper

https://algorithmwatch.org/de/automatisierte-diskriminierung-facebook-verwendet-grobe-stereotypen-um-die-anzeigenschaltung-zu-optimieren/

➔ Bias (und mögliche Diskriminierung) durch

- Daten
  - Selektionsbias (wurde hier beleuchtet, es gibt natürlich weitere Risiken)
  - Ein Ansatz: repräsentativer sampeln
- Repräsentation
  - Reporting bias and social desirability bias; Normalisierung gesellschaftlicher Machtverhältnisse
  - Ein Ansatz: Stakeholder-Einbindung / Co-Design in Prozessen der Wissensgenerierung
- Entscheidungen
  - Vorurteile u/o “rational discrimination”
  - Ein Ansatz: De-Biasing von Algorithmen
  - Ein anderer Ansatz: Verfahrenstechnische Schutzmaßnahmen , Verfahrensgarantien (procedural safeguards)

Kumulation

Eine Entscheidung kommt selten allein

Frauen erhalten von Algorithmen Stellenausschreibungen für andere, häufig schlechter bezahlte Jobs

Eine Entscheidung kommt selten allein

Menschen nehmen eine Welt wahr,
In der eher Männer gut bezahlte Jobs haben
(Beobachtung u/o medial/textuell vermittelt) *

Es gibt weniger Frauen in gut bezahlten Jobs

Frauen erhalten von Algorithmen Stellenausschreibungen für andere, häufig schlechter bezahlte Jobs

*) Mögliche Verstärkung der Effekte bei Verwendung von *pre-trained language models*, vgl Delobelle, Tokpo, Calders, & Berendt:, NAACL 2022; to appear, 2022; Weitere Referenzen in: Berendt, in Rostalski & BMU 2022).

Eine Entscheidung kommt selten allein

Menschen nehmen eine Welt wahr,
In der eher Männer gut bezahlte Jobs haben
(Beobachtung u/o medial/textuell vermittelt) *

Es gibt weniger Frauen in gut bezahlten Jobs

Frauen erhalten von Algorithmen Stellenausschreibungen für andere, häufig schlechter bezahlte Jobs

Ausschreibungen werden so formuliert, dass sich Frauen tendenziell seltener auf gut bezahlte Jobs bewerben

*) Mögliche Verstärkung der Effekte bei Verwendung von *pre-trained language models*, vgl Delobelle, Tokpo, Calders, & Berendt:, NAACL 2022; to appear, 2022; Weitere Referenzen in: Berendt, in Rostalski & BMU 2022).

Auch in “digital-lastigen” Umgebungen werden viele Entscheidungen von Menschen getroffen

Wörter in (auch geschlechtsneutral formulierten) Stellenanzeigen	
eigenständig, selbstständig	Verantwortung übernehmen
ambitioniert, leistungsfähig	committed, engagiert
zielorientiert	zuverlässig
direkt, bestimmt	ehrlich

https://www.spiegel.de/start/stellenanzeigen-werden-oft-fuer-maenner-formuliert-wie-frauen-trotzdem-den-job-bekommen-a-2ff0215c-009c-48b1-b045-a462ca808cd7

Auch in “digital-lastigen” Umgebungen werden viele Entscheidungen von Menschen getroffen

Agentisch	Kommunal
eigenständig, selbstständig	Verantwortung übernehmen
ambitioniert, leistungsfähig	committed, engagiert
zielorientiert	zuverlässig
direkt, bestimmt	ehrlich

- Signifikant weniger Bewerbungen von Frauen auf Stellenanzeigen, die zu viele „agentische“ Wörter enthalten
- Anpassung des Wortprofils zu „kommunal“ ➔ bis zu 33 Prozent mehr Bewerbungen von Frauen
- Kein entsprechender Effekt bei Männern – sie bewerben sich eh.

https://www.spiegel.de/start/stellenanzeigen-werden-oft-fuer-maenner-formuliert-wie-frauen-trotzdem-den-job-bekommen-a-2ff0215c-009c-48b1-b045-a462ca808cd7

Eine Entscheidung kommt selten allein

Menschen nehmen eine Welt wahr,
In der eher Männer gut bezahlte Jobs haben
(Beobachtung u/o medial/textuell vermittelt) *

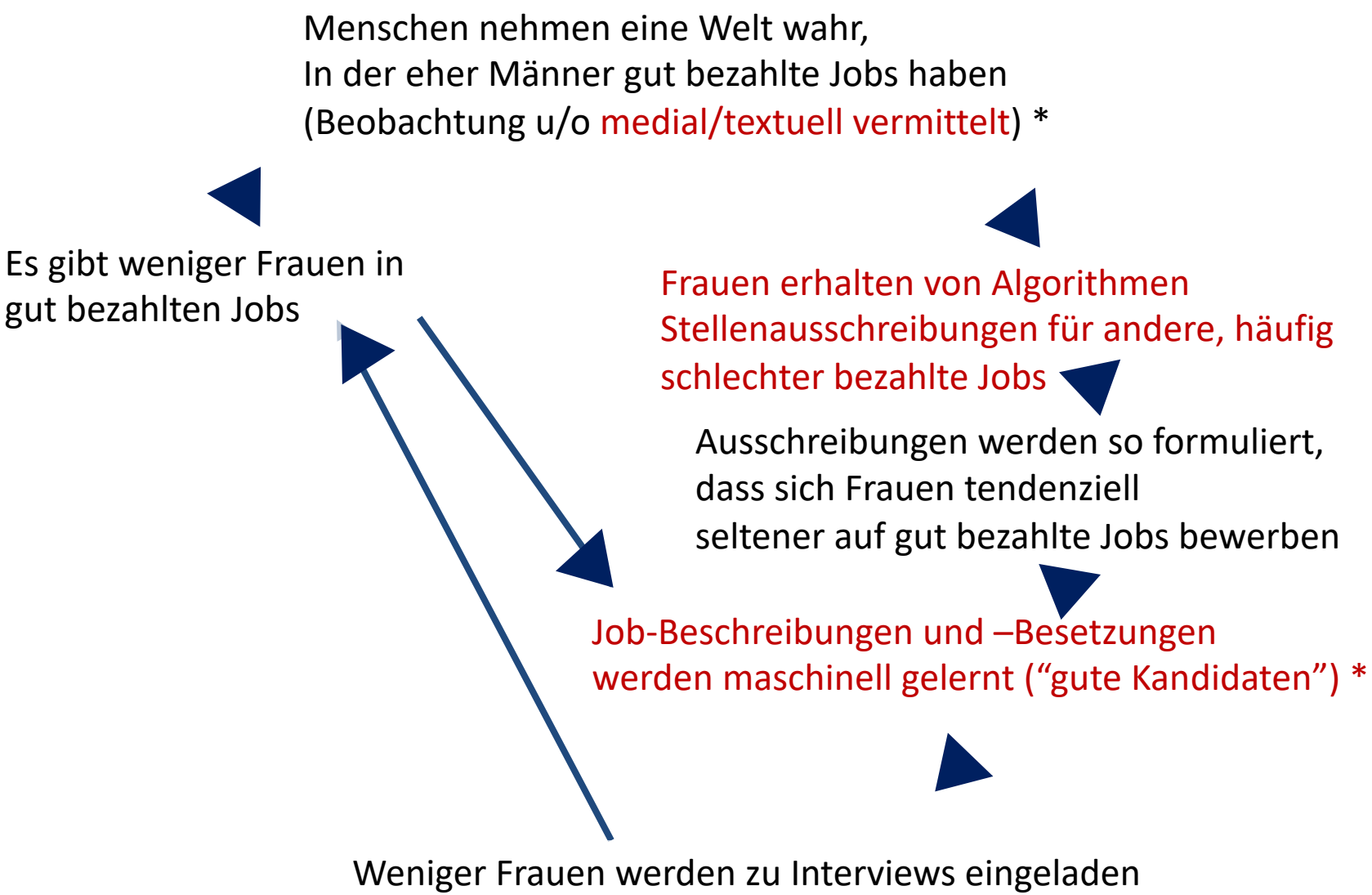
Es gibt weniger Frauen in gut bezahlten Jobs

Frauen erhalten von Algorithmen Stellenausschreibungen für andere, häufig schlechter bezahlte Jobs

Ausschreibungen werden so formuliert, dass sich Frauen tendenziell seltener auf gut bezahlte Jobs bewerben

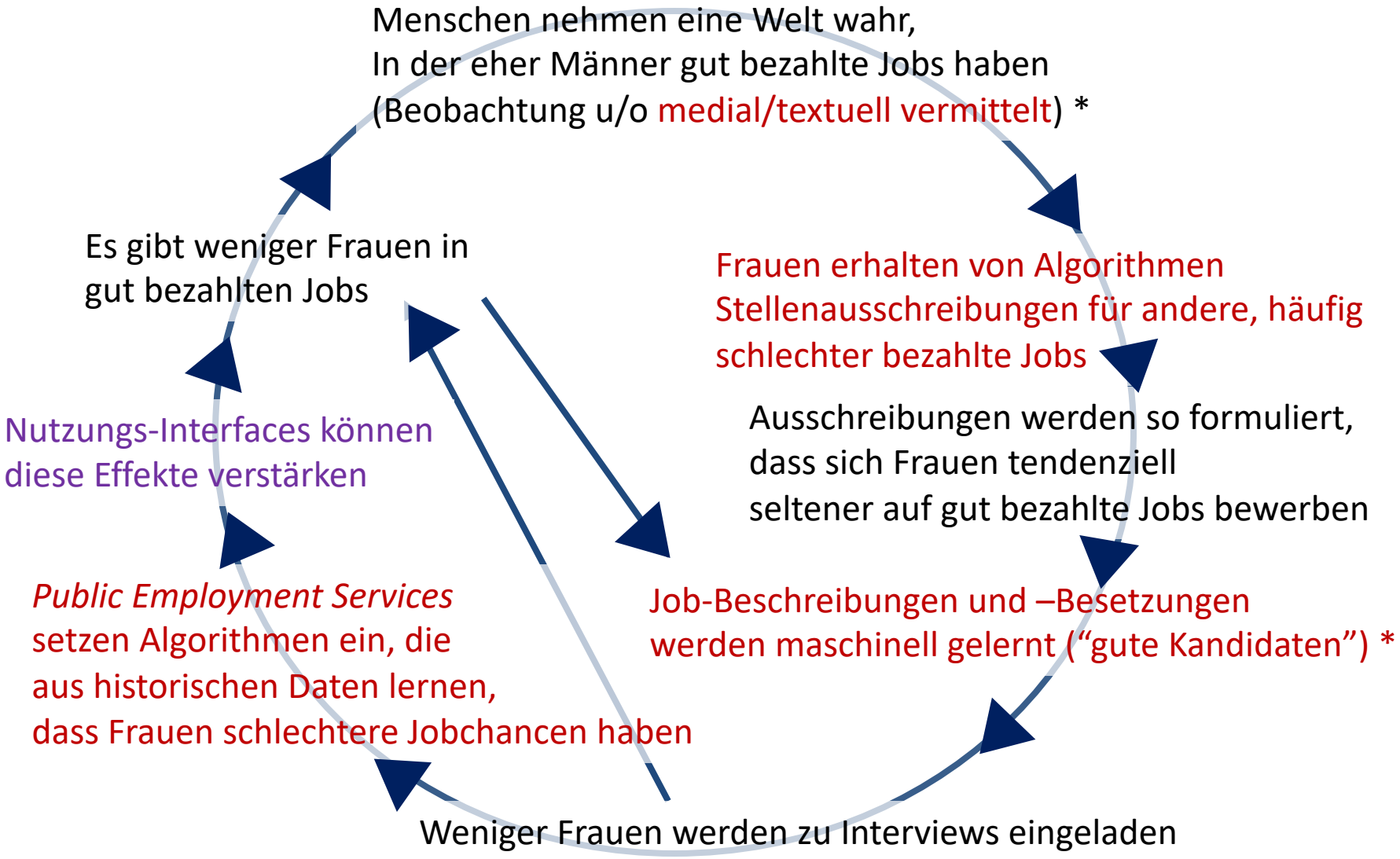
*) Mögliche Verstärkung der Effekte bei Verwendung von *pre-trained language models*, vgl Delobelle, Tokpo, Calders, & Berendt:, NAACL 2022; to appear, 2022; Weitere Referenzen in: Berendt, in Rostalski & BMU 2022).

Eine Entscheidung kommt selten allein



*) Mögliche Verstärkung der Effekte bei Verwendung von *pre-trained language models*, vgl Delobelle, Tokpo, Calders, & Berendt:, NAACL 2022; to appear, 2022;
Weitere Referenzen in: Berendt, in Rostalski & BMU 2022).

Eine Entscheidung kommt selten allein



*) Mögliche Verstärkung der Effekte bei Verwendung von *pre-trained language models*, vgl Delobelle, Tokpo, Calders, & Berendt:, NAACL 2022; to appear, 2022; Weitere Referenzen in: Berendt, in Rostalski & BMU 2022).

➔ Bias (und mögliche Diskriminierung) durch

- Daten
  - Selektionsbias (wurde hier beleuchtet, es gibt natürlich weitere Risiken)
  - Ein Ansatz: repräsentativer sampeln
- Repräsentation
  - Reporting bias and social desirability bias; Normalisierung gesellschaftlicher Machtverhältnisse
  - Ein Ansatz: Stakeholder-Einbindung / Co-Design in Prozessen der Wissensgenerierung
- Entscheidungen
  - Vorurteile u/o “rational discrimination”
  - Ein Ansatz: De-Biasing einzelner Schritte
  - Ein Ansatz: Verfahrenstechnische Schutzmaßnahmen , Verfahrensgarantien (procedural safeguards)
- Kumulation
  - “pernicious feedback loops” lassen einzelne Entscheidungen risikoarm u/o rational erscheinen
  - Ein Ansatz: systemische Analyse

”Sei doch nicht immer so negativ”

(vgl. Andreas Bischofs Punkte)

Zwei gute Waffen

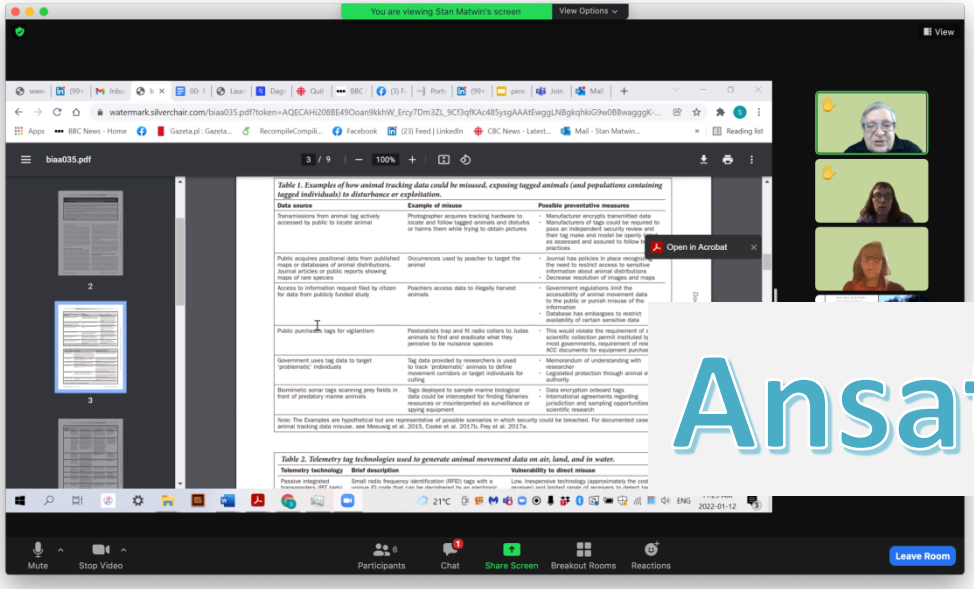
- Datenschutzprinzipien und -recht
- “Selbst denken”: z.B. AI Snake Oil (Narayanan, 2019, 2021)

Aktuelle Lehrveranstaltungen!

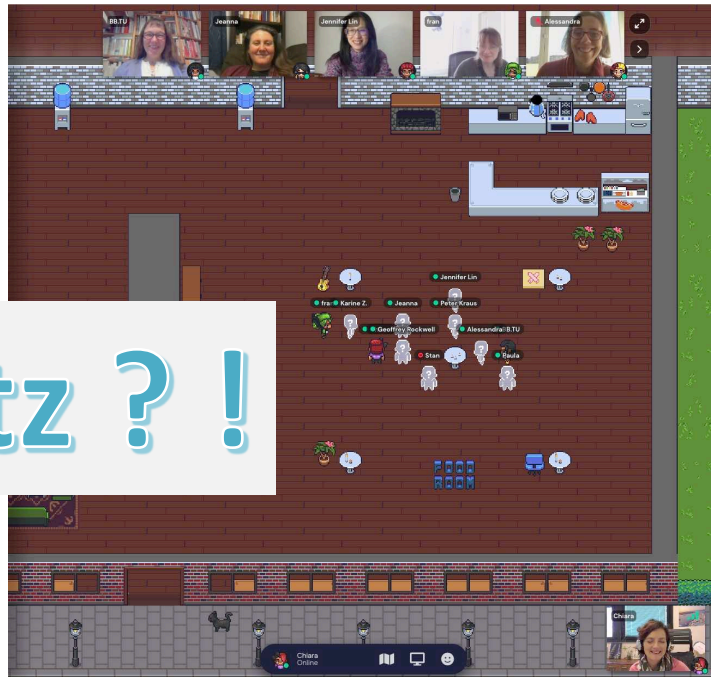
Von Technikfolgenabschätzung zur Nutzung emanzipatorischer Potenziale

Aktuelles, nicht fiktives Beispiel:

mittelbare Diskriminierung durch Gleichbehandlung in akademischen Job-Interviews: “nur in Präsenz”



Ansatz ? !



Berendt et al. (Dagstuhl Reports 2022)

u.a.


Wie wollen wir digital leben?
VORTRÄGE 1
AUSSTELLUNG 2
PERFORMANCE 3
ANMELDUNG ERBETEN

Zitierte Quellen

Berendt, B. (2022). The AI Act Proposal: Towards the next transparency fallacy? Why AI regulation should be based on principles based on how algorithmic discrimination works. To appear in F. Rostalski & BMU (Eds.), Künstliche Intelligenz - Wie gelingt eine vertrauenswürdige Verwendung in Deutschland und Europa? Tübingen, Germany: Mohr Siebeck. (last author version)

Bettina Berendt, Stan Matwin, Chiara Renso, Fran Meissner, Francesca Pratesi, Alessandra Raffaetà, Geoffrey Rockwell: Mobility Data Mining: from Technical to Ethical (Dagstuhl Seminar 22022). Dagstuhl Reports 12(1): 35-66 (2022)

Joy Buolamwini, Timnit Gebru: Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. FAT 2018: 77-91

Pieter Delobelle, Ewoenam Kwaku Tokpo, Toon Calders, Bettina Berendt:
Measuring Fairness with Biased Rulers: A Comparative Study on Bias Metrics for Pre-trained Language Models. NAACL-HLT 2022: 1693-1706

Friedman, Batya and Nissenbaum, Helen, Bias in computer systems, ACM Transactions on Information Systems, 14(3), 1996, 330-347

Milagros Miceli, Julian Posada, Tianling Yang: Studying Up Machine Learning Data: Why Talk About Bias When We Mean Power? Proc. ACM Hum. Comput. Interact. 6(GROUP): 1-14 (2022)

Milagros Miceli, Martin Schuessler, Tianling Yang: Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision. Proc. ACM Hum. Comput. Interact. 4(CSCW2): 115:1-115:25 (2020)

Arvind Narayanan's AI Snake Oil presentation (slides) and interview (in the Introduction to the FakeAIBook);

Kristen M. Scott, Sonja Mei Wang, Milagros Miceli, Pieter Delobelle, Karolina Sztandar-Sztanderska, Bettina Berendt: Algorithmic Tools in Public Employment Services: Towards a Jobseeker-Centric Perspective. FAccT 2022: 2138-2148

Williams, A. Miceli, M., & Gebru, T. (2022). The exploited labor behind artificial intelligence. Noema Magazine 2022. https://www.noemamag.com/the-exploited-labor-behind-artificial-intelligence/

Lehrveranstaltungen Überblick: https://people.cs.kuleuven.be/~bettina.berendt/teaching/