UNIVERSITÄT DES SAARLANDES

Wenn Algorithmen über die Zukunft entscheiden

Eine 360° Betrachtung der Erwartungen an Algorithmen im Kontext Fairness und Diskriminierung

Dr. Markus Langer – Universität des Saarlandes

Algorithmische Entscheidungen beeinflussen unser Leben

Millions of black people affected by racial bias in health-care algorithms

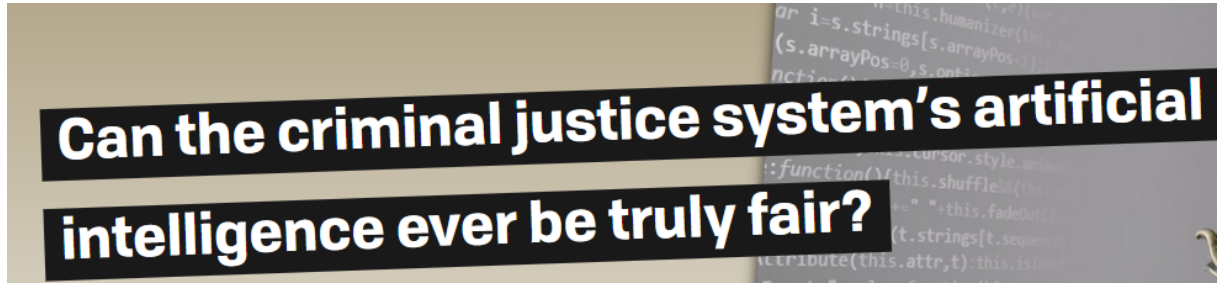
hiips://www.nature.com/articles/d41586-019-03228-6

Amazon scraps secret AI recruiting tool that showed bias against women

hiips://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Coded Bias review: Eye-opening Netflix doc faces racist technology

hiips://www.cnet.com/culture/entertainment/coded-bias-review-eye-opening-netflix-documentary-faces-up-to-racist-tech/



hiips://massivesci.com/articles/machine-learning-compas-racism-policing-fairness/

Perspektiven auf Bias und Diskriminierung in algorithmischen Entscheidungen

UNIVERSITÄT DES SAARLANDES

- Wie funktioniert algorithmische Entscheidungsfindung?
- Die Perspektive (mancher) Anbieter
- Die ethische und rechtliche Perspektive
- Perspektive der Betroffenen + Studienergebnisse
- Perspektive der Nutzer + Studienergebnisse

Perspektiven auf Bias und Diskriminierung in algorithmischen Entscheidungen

UNIVERSITÄT DES SAARLANDES

- Wie funktioniert algorithmische Entscheidungsfindung?
- Die Perspektive (mancher) Anbieter
- Die ethische und rechtliche Perspektive
- Perspektive der Betroffenen + Studienergebnisse
- Perspektive der Nutzer + Studienergebnisse

Was ist ein Algorithmus?

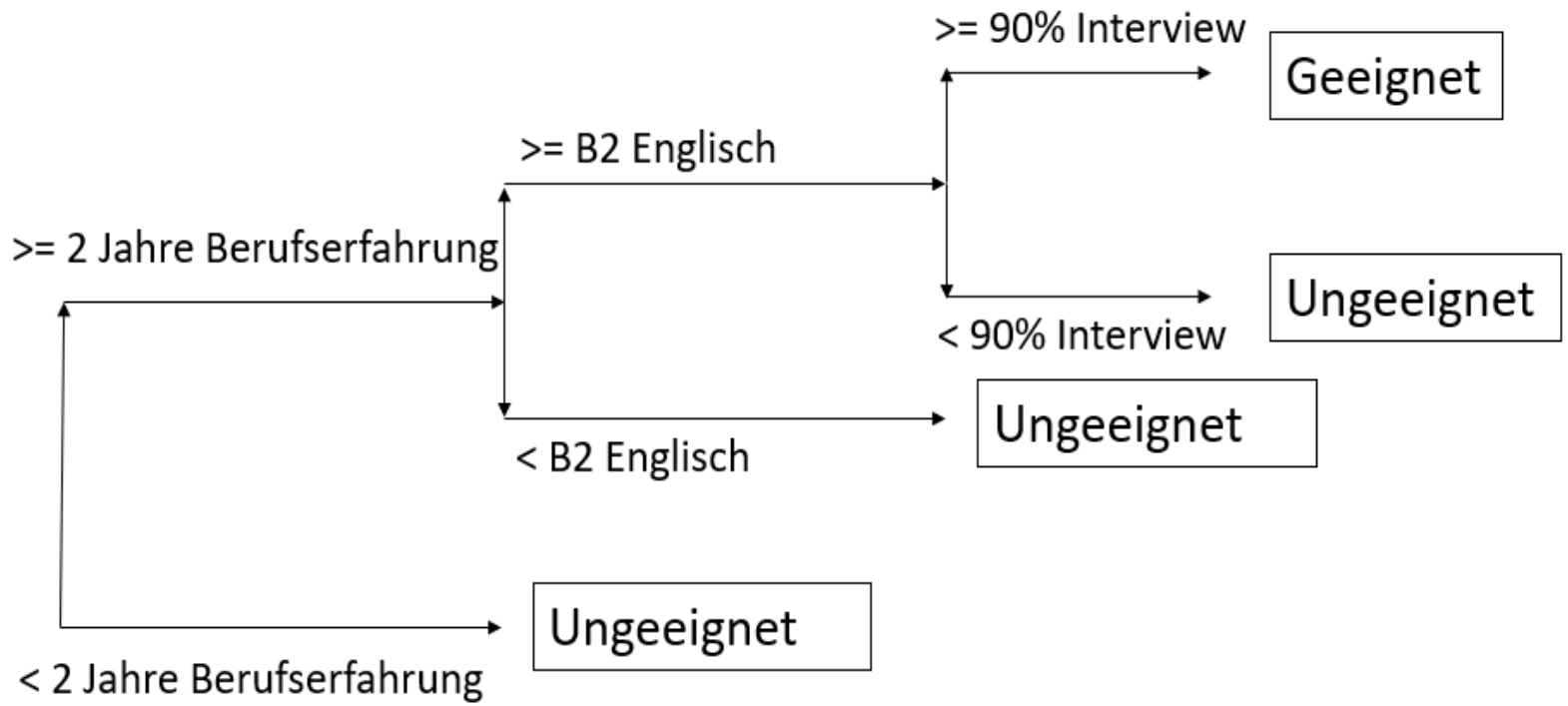
- Eine Vorgehensweise, um ein Problem zu lösen; ein Rechenvorgang nach einem bestimmten Schema (Duden)

→Algorithmen nehmen Inputs (z.B. Prädiktoren oder Features), verarbeiten diese (z.B. mittels statistischer Verfahren), geben Outputs aus (z.B. Klassifikationen, Vorhersagen, Bewertungen…)

- Beispiel für einen Algorithmus: Body-Mass-Index (BMI)
  - Input: Gewicht + Körpergröße
  - Formel: Gewicht / Körpergröße in Quadrat
  - Output: Body-Mass-Index

Arten von Algorithmen

- Händisch programmiert, basierend auf Expertenwissen



Arten von Algorithmen

- Algorithmen basierend auf Verfahren des maschinellen Lernens

hiips://de.wikipedia.org/wiki/Random_Forest

Arten von Algorithmen

hiips://www.fz-juelich.de/portal/DE/Presse/beitraege/2018/ki-im-dienste-der-wissenschaft/_node.html

Entwicklungsprozess eines Algorithmus

- Wichtig ist, dass diese Schritte in **getrennten Datensätzen passieren**

  (z.B. Trainings-, Validierungs-, Testdatensatz) Beispiel:

  - 80% der **vorhandenen Daten** fürs Trainieren
  - 20% der **vorhanden Daten** fürs Validieren
  - **Unabhängige, neue Daten** fürs Testen
- Denn das Ziel ist: **Generalisierbarkeit des Algorithmus**

Entwicklungsprozess eines Algorithmus

- Beispiel: Automatisierte Bewertung von Bewerbungsgesprächen
- Kontext: Asynchrone Videointerviews

Entwicklungsprozess eines Algorithmus

- Wir haben 1000 bereits bewertete Interviews
  - Wir wissen, welche der Bewerber gut oder weniger gut abgeschnitten haben
- Wir nutzen 800 dieser Interviews, um den Algorithmus zu „trainieren“
  - Automatisiert trennt der Algorithmus zwischen guten und weniger guten Bewerber
- Wir nutzen 200 Interviews, die der Algorithmus noch nicht „gesehen“ hat, um zu testen, wie gut unser Algorithmus funktioniert
  - Wir sehen, wie gut der Algorithmus bei den ungesehenen Bewerbern zwischen guten und weniger guten Bewerbern trennen kann

Beispielhaftes Ziel: Klassifikation

Grün = geeignete Bewerber
Blau = ungeeignete Bewerber

hiips://data-science-blog.com/blog/2017/12/20/maschinelles-lernen-klassifikation-vs-regression/

Entwicklungsprozess eines Algorithmus

- Wir können den Algorithmus nun auf neue Bewerbungsgesprächsvideos anwenden und diese automatisiert bewerten lassen
- Wichtig: Validierung, kontinuierliche Überprüfung, Monitoring, Updating

Perspektiven auf Bias und Diskriminierung in algorithmischen Entscheidungen

- Wie funktioniert algorithmische Entscheidungsfindung?
- Die Perspektive (mancher) Anbieter
- Die ethische und rechtliche Perspektive
- Perspektive der Betroffenen + Studienergebnisse
- Perspektive der Nutzer + Studienergebnisse

Die Perspektive (mancher) Anbieter

(Raghavan et al., 2020)

Vendor	Claim about bias
HireVue	Provide "a highly valid, bias-mitigated assessment"
pymetrics	"…the Pre-Hire assessment does not show bias against women or minority respondents."
PredictiveHire	"AI bias is testable, hence fixable."
Knockri	"Knockri's A.I. is unbiased because of its full spectrum database that ensures there's no benchmark of what the 'ideal candidate' looks like."

Table 2: Examples of claims that vendors make about bias, taken from their websites.

Die Perspektive (mancher) Anbieter:

Algorithmische Systeme sind weniger biased als Menschen ODER der Bias ist kontrollierbarer

Perspektiven auf Bias und Diskriminierung in algorithmischen Entscheidungen

- Wie funktioniert algorithmische Entscheidungsfindung?
- Die Perspektive (mancher) Anbieter
- Die ethische und rechtliche Perspektive
- Perspektive der Betroffenen + Studienergebnisse
- Perspektive der Nutzer + Studienergebnisse

Die ethische und rechtliche Perspektive (Jobin et al., 2019)

Table 3 | Ethical principles identified in existing AI guidelines

Ethical principle	Number of documents	Included codes
Transparency	73/84	Transparency, explainability, explicability, understandability, interpretability, communication, disclosure, showing
Justice and fairness	68/84	Justice, fairness, consistency, inclusion, equality, equity, (non-) bias, (non-)discrimination, diversity, plurality, accessibility, reversibility, remedy, redress, challenge, access and distribution
Non-maleficence	60/84	Non-maleficence, security, safety, harm, protection, precaution, prevention, integrity (bodily or mental), non-subversion
Responsibility	60/84	Responsibility, accountability, liability, acting with integrity

Auszug aus dem „AI Act“ der EU

KI-Systeme, die zu diesem Zweck eingesetzt werden, können zur Diskriminierung von Personen oder Gruppen führen und historische Diskriminierungsmuster, beispielsweise aufgrund der rassischen oder ethnischen Herkunft, einer Behinderung, des Alters oder der sexuellen Ausrichtung, fortschreiben oder neue Formen von Diskriminierung mit sich bringen

hiips://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX%3A52021PC0206

Die ethische und rechtliche Perspektive

Algorithmische Systeme können Biases unterliegen – Diese Biases müssen verhindert, kontrolliert, reguliert werden

Perspektiven auf Bias und Diskriminierung in algorithmischen Entscheidungen

- Wie funktioniert algorithmische Entscheidungsfindung?
- Die Perspektive (mancher) Anbieter
- Die ethische und rechtliche Perspektive
- Perspektive der Betroffenen + Studienergebnisse
- Perspektive der Nutzer + Studienergebnisse

Studie an Betroffenen – Methoden (Schlicker et al., 2021)

- Online Studie (N = 209 medizinisches Fachpersonal)
  - Teilnehmende stellten sich vor Urlaub zu beantragen
  - Eine weitere Person will zur gleichen Zeit Urlaub
  - Die Beiden können sich nicht einigen, wer Urlaub nehmen darf
  - Menschlicher Entscheidungsträger vs. Algorithmische Entscheidung
- Maße
  - Wahrgenommene Gerechtigkeit

Ergebnisse + weitere Studien

UNIVERSITÄT DES SAARLANDES

- Algorithmus wird als **konsistenter** wahrgenommen aber Teilnehmerinnen erwarteten, dass sie **weniger Einfluss** auf das Ergebnis haben, es weniger kontrollieren können

- Weitere Studien unterstützen Konsistenzerwartungen bezüglich algorithmischer Entscheidungssysteme (siehe Langer et al., 2019, Lee, 2018, für einen Überblick siehe Langer & Landers, 2021)

Perspektive der Betroffenen – Studie II

- Betroffene teils negativ eingestellt hinsichtlich (vollständig) automatisierter Entscheidungen (Langer & Landers, 2021)
  - Besonders in automatisierten Bewerbungsgesprächen
- Offensichtliche Gründe: weniger zwischenmenschlicher Kontakt, zeigt weniger „Anstrengung“ von Seiten des Unternehmens
- Weniger offensichtlicher Grund: Bewerberinnen bevorzugen menschliche „Zufälligkeit“ oder „Flexibilität“ (Dietvorst et al., 2020)
  - **Menschliche Entscheidungen**: Ich hoffe Glück zu haben; hoffe auf Bias zu meinen Gunsten
  - **Algorithmische Entscheidungen**: Wenn ich weniger geeignet bin, werde ich abgelehnt; Wenn ich die Mindestanforderungen nicht erfülle, werde ich abgelehnt

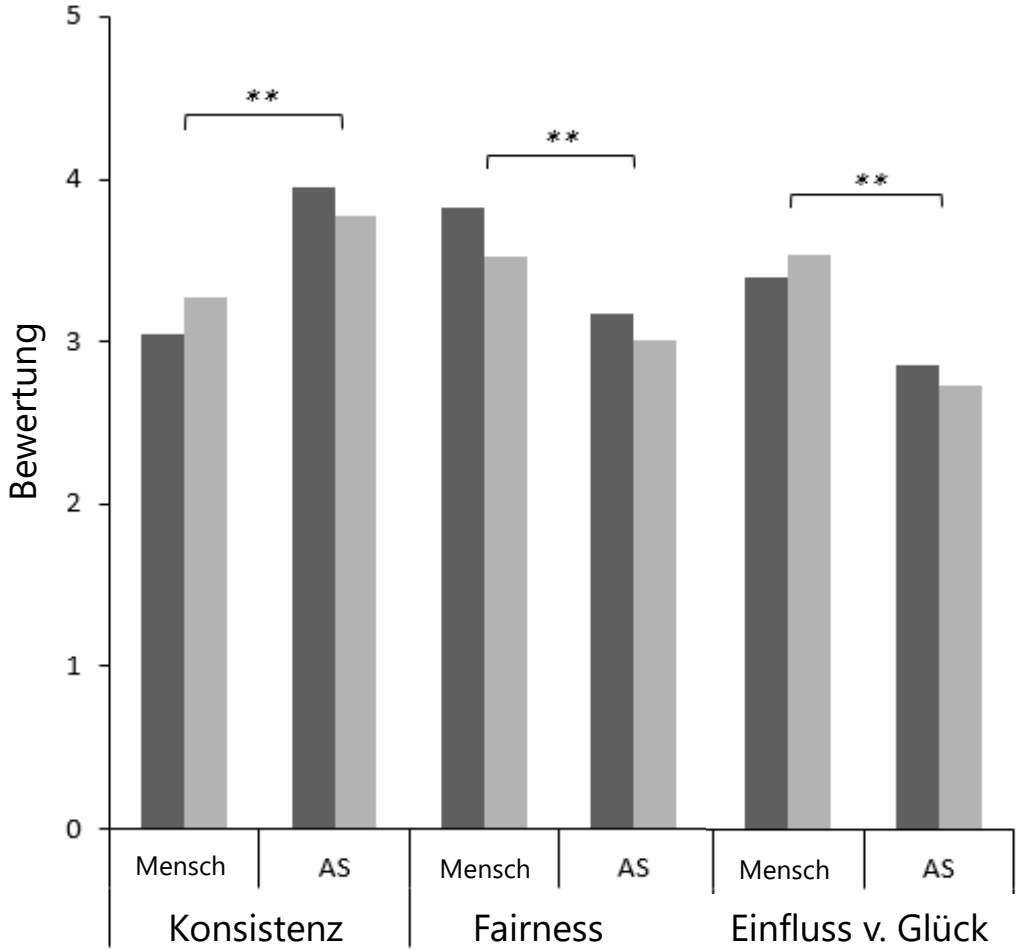
Perspektive der Betroffene – Studie II, Methoden

(Langer et al., 2021)

- Online Studie (N = 150 Studenten)
  - Teilnehmer stellten sich vor, sie bewerben sich auf einen Job
  - Teilnehmer befanden sich in der Rolle eines geeigneten vs. ungeeigneten Bewerbers
  - Mensch vs. Algorithmus evaluiert Bewerbungsgespräch

- Maße
  - Wahrgenommene Fairness
  - Wahrgenommener Einfluss verschiedener Bewerbercharakteristiken auf die Eignungsbewertung

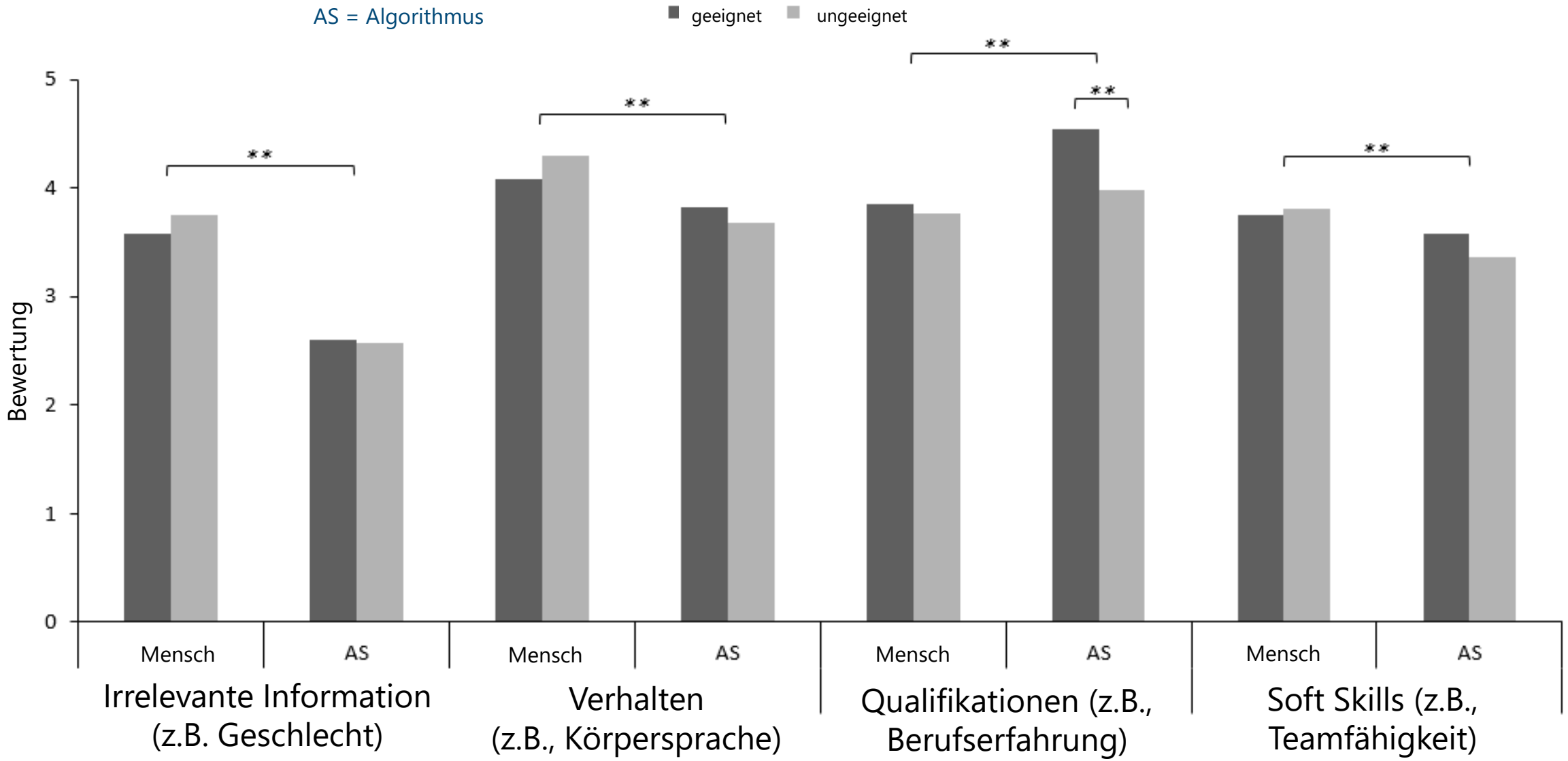
Perspektive der Betroffene – Ergebnisse



AS = Algorithmus

Perspektive der Betroffene – Wahrgenommener Einfluss verschiedener Informationen

AS = Algorithmus



Perspektive der Betroffenen - Zusammenfassung

- Algorithmische Entscheidungen werden als **konsistenter**, aber weniger beeinflussbar wahrgenommen
- Wahrnehmung, dass Algorithmen mehr auf „harten Fakten“ und Menschen mehr auf Soft Skills und irrelevante Faktoren achten
- Aber: Positivere Wahrnehmung menschlicher Entscheidungen

Perspektive der Betroffenen

Algorithmische Systeme unterliegen weniger Bias – Menschen bevorzugen aber dennoch Menschen, evtl. weil sie einfacher zu beeinflussen sind

Perspektiven auf Bias und Diskriminierung in algorithmischen Entscheidungen

- Wie funktioniert algorithmische Entscheidungsfindung?
- Die Perspektive (mancher) Anbieter
- Die ethische und rechtliche Perspektive
- Perspektive der Betroffenen + Studienergebnisse
- Perspektive der Nutzer + Studienergebnisse

Die Perspektive der Nutzer - Beispielstudie

- Vertrauen als Beispielperspektive (Lee & See, 2004, Mayer et al., 1995)
- Vertraue ich einem algorithmischen System…
  - … eine Aufgabe erfolgreich zu bearbeiten und meinen Werten und Moralvorstellungen gerecht zu werden? (Hoff et al., 2015, Lee & See, 2004)
  - …obwohl ich nicht weiß, wie es funktioniert, und obwohl es möglicherweise Biases und fehlerhafte Outputs geben kann?
- Zu hohes sowie zu niedriges Vertrauen als Problem (Parasuraman et al., 2000)
  - Zu hohes Vertrauen: Fehler werden übersehen, blindes Vertrauen
  - Zu niedriges Vertrauen: Ineffizient (ständige Kontrolle)

Klassische Befunde zu Vertrauen in automatisierte Systeme
(in klassischen Bereichen der Automatisierung)

- Anfängliches Vertrauen in automatisierte Systeme ist hoch – evtl. sogar höher als gegenüber Menschen
- Vertrauensbrüche lassen das Vertrauen stark abfallen – teils stärker als bei Menschen
- Vertrauen wiederaufzubauen scheint schwieriger als bei Menschen (DeVisser et al., 2016; Madhaven et al., 2007)
- Warum?
  - Erwartung der Perfektion und Konsistenz bei automatisierten Systemen?
  - Erwartung der Imperfektion und Flexibilität bei Menschen?

hiips://www.flaticon.com/de/kostenloses-icon/radar_196345

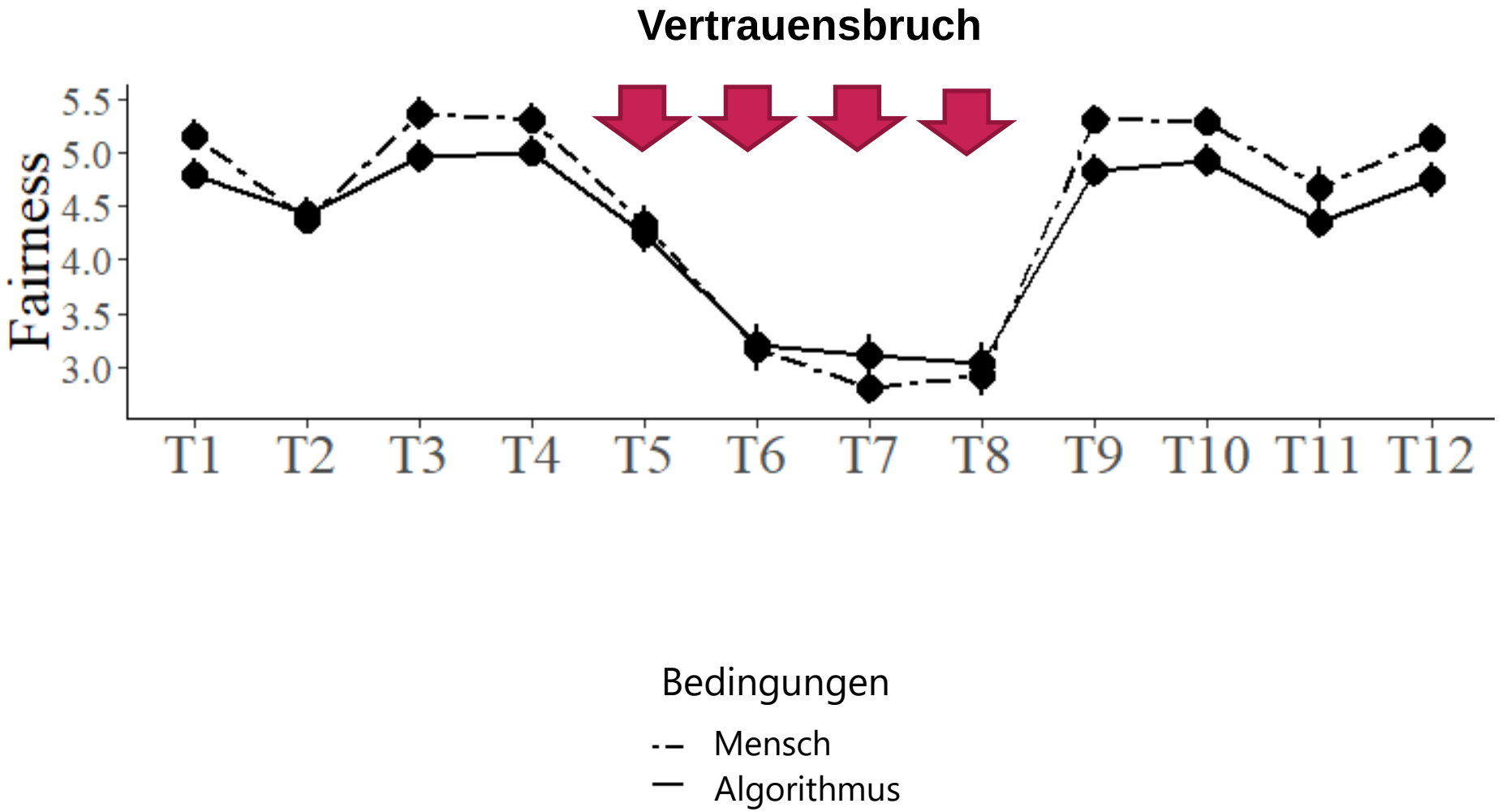
Was ist mit Kontexten in denen Biases eine Rolle spielen könnten?

- Bisherige Forschung fokussiert sich auf klassische Bereiche der Automatisierung
  - Vertrauensbrüche basierend auf „harten“ Leistungskriterien

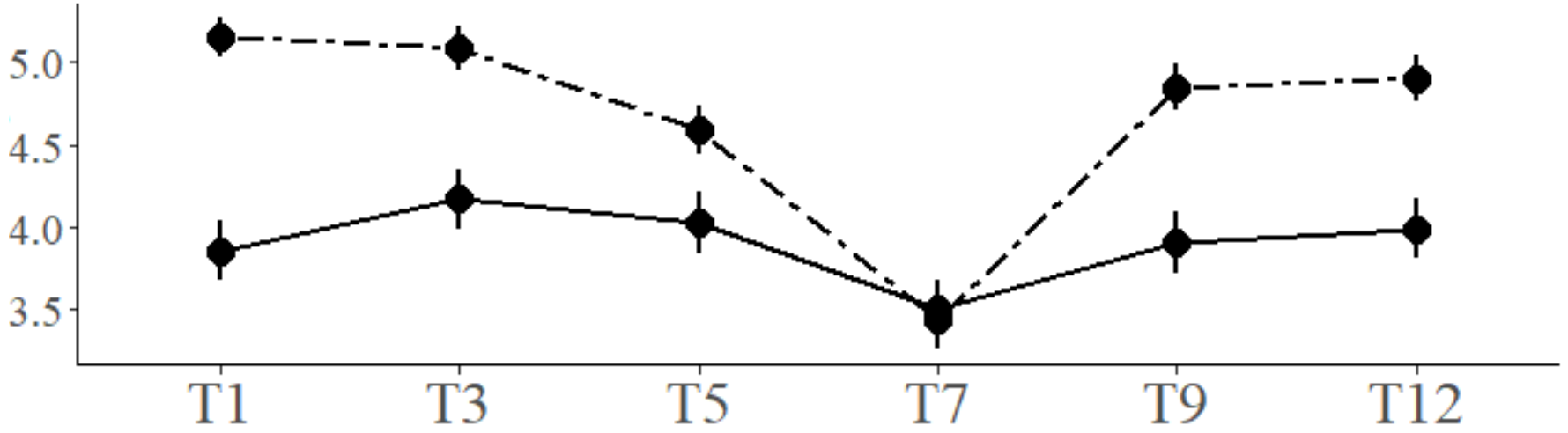
- Für viele Einsatzbereiche algorithmischer Entscheidungen, werden weitere Leistungskriterien wichtig
  - Entsprechen Systemoutputs ethischen und moralischen Standards?
  - Sind Outputs verständlich und lassen sie sich rechtfertigen?

Studie mit Nutzern - Methoden

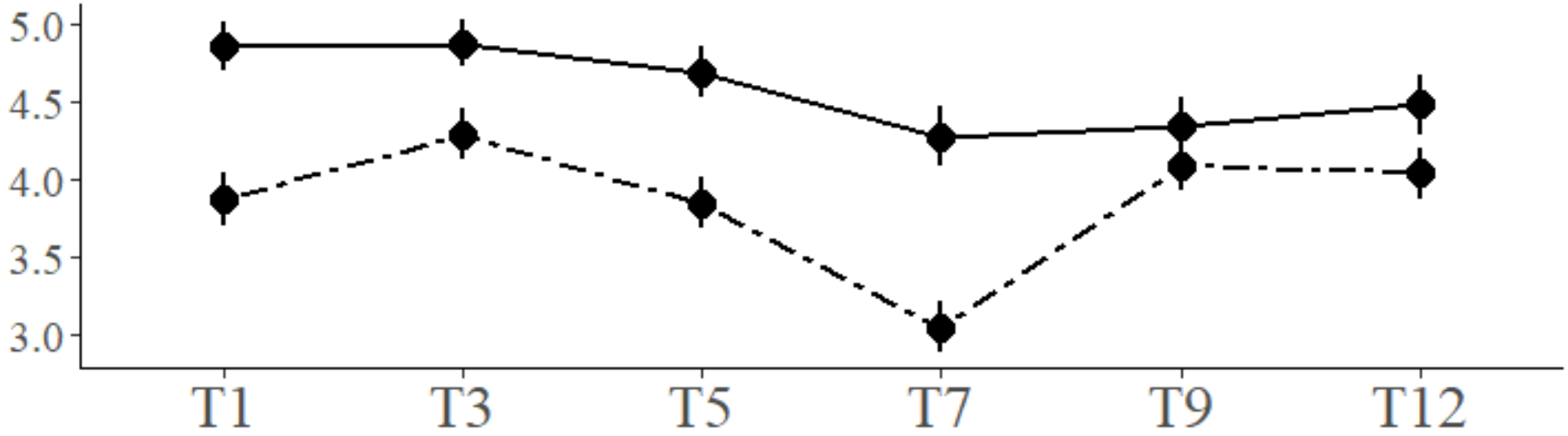
- Online Studie; Teilnehmende (N = 121)
- Kontext: Entscheidungsfindung in der Personalauswahl
- 12 Runden Personalauswahlaufgaben:
  - Entscheidungshilfe: Algorithmus vs. Mensch
  - Vertrauensbruch (d.h. vorwiegend männliche Vorauswahl in Runden 5.-8.)
  - Nach der 8. Runde wieder ausbalancierte Vorauswahl (d.h. ähnlich viele Frauen wie Männer)


Vertrauensbruch
Fairness
5.5
5.0
4.5
4.0
3.5
3.0
T1
T2
T3
T4
T5
T6
T7
T8
T9
T10
T11
T12
Bedingungen
Mensch
Algorithmus

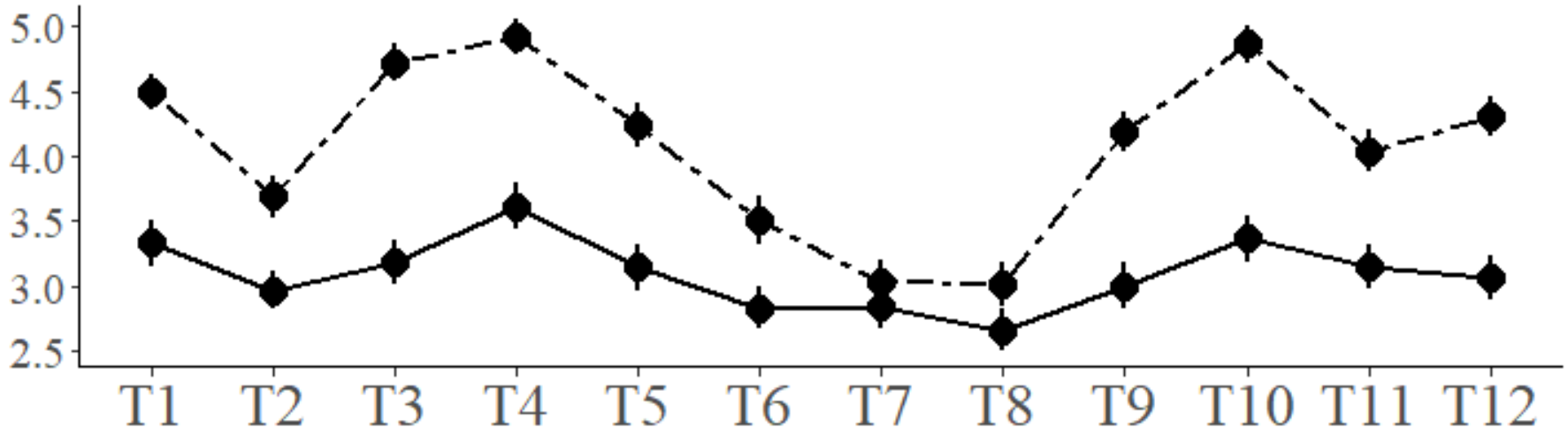
Fähigkeit



Unvoreingenommenheit



Vertrauen



Bedingungen

-- Mensch

— Algorithmus

Perspektive der Nutzer - Zusammenfassung

- Konsistenzannahme? Neigen Nutzer dazu, zu denken, dass Algorithmen eher weniger Biases produzieren?
  - Weitere Studien deuten darauf hin, dass Menschen weniger stark auf Biases reagieren, weil Algorithmen keine „Intention“ zur Diskriminierung haben (Bigman et al., 2020)

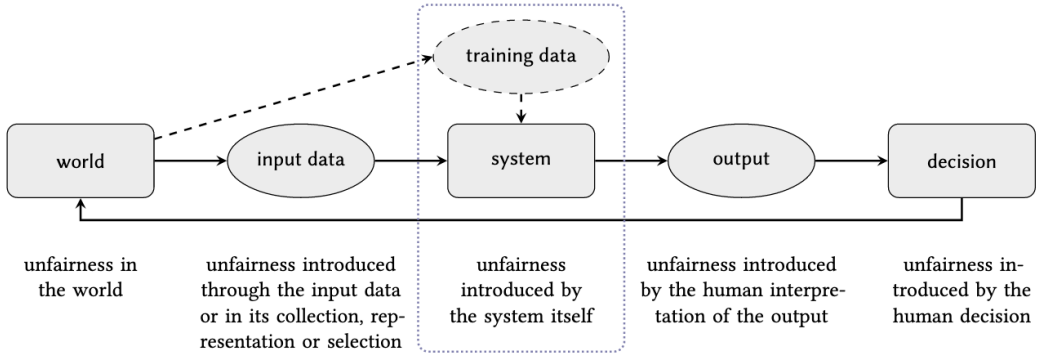
Perspektive der Nutzer
Algorithmen sind konsistent, geringere Erwartung von Biases im Vergleich zu menschlichen Entscheidungen

Allgemeine Zusammenfassung

- (Manche) **Anbieter** behaupten, dass Algorithmen weniger Biases produzieren als Menschen oder diese Biases kontrollierbarer sind
- **Aus ethischer und rechtlicher Perspektive** ist das Besorgnis groß über mögliche algorithmische Diskriminierung; es wird betont, dass Algorithmen Biases produzieren können
- **Betroffene Personen** erwarten, dass Algorithmen konsistent aber auch weniger beeinflussbar sind als Menschen; manchmal hoffen Betroffene sogar auf menschliche Biases
- **Nutzer** erwarten (anfänglich?), dass Algorithmen konsistent sind; reagieren teils weniger stark auf mögliche Biases in den Outputs von Algorithmen

Wie entsteht möglicher Bias in algorithmischen Systemen? (z.B. Tay et al., 2021)

- Durch Ungleichheiten in der Welt
- Durch die Auswahl von Trainingsdaten
- Durch die Auswahl von Inputdaten
- Durch die Wahl der Algorithmusart
- Durch die Interpretation von Outputs durch Menschen
- ....



Fazit I

- Eine Erwartung von Nutzern und betroffenen Personen scheint, das **Algorithmen konsistenter sind und weniger Biases unterliegen als Menschen**

- Argumente von Anbietern, dass Algorithmen weniger Biases zeigen, könnten Nutzer, Betroffene, Entscheiderinnen in Unternehmen überzeugen – **unfaire Biases werden nicht erwartet, evtl. weniger wahrscheinlich aufgedeckt**

- Falls unfaire Biases aufgedeckt werden→ **starke negative Reaktionen**, negative Publicity für Anbieter und für Unternehmen, die solche Systeme nutzen; **Verlust öffentlichen Vertrauens** in algorithmische Entscheidungsfindung

Fazit II

Algorithmische Systeme in der Entscheidungsfindung **haben durchaus das Potenzial Biases zu reduzieren** – aber das bleibt eine Herausforderung. Es wird wahrscheinlicher, wenn…

- …algorithmische Systeme **gut designed** sind, kontinuierlich **überprüft und upgedated** werden
- …Personen, die diese Systeme nutzen und kontrollieren, **gut trainiert** sind
- …Personen, die diese Systeme nutzen und kontrollieren, **adäquate Informationen** über und durch das System erhalten
- **…alle Interessensgruppen** eine **kritische Perspektive** bezüglich möglicher algorithmischer Biases und Diskriminierung einnehmen

Literaturempfehlungen

- **Qualitätskriterien für algorithmische Diagnostik:** Landers, R. N., & Behrend, T. S. (2022). Auditing the AI auditors: A framework for evaluating fairness and bias in high stakes AI predictive models. *American Psychologist*.
- **Übersichtsarbeit zu Reaktionen von Betroffenen auf algorithmische Entscheidungen:** Langer, M., & Landers, R. N. (2021). The future of artificial intelligence at work: A review on effects of decision automation and augmentation on workers targeted by algorithms and third-party observers. *Computers in Human Behavior*, *123*, 106878.
- **Übersichtsarbeit zur Entstehung möglicher Biases in algorithmischen Entscheidungen:** Tay, L., Woo, S. E., Hickman, L., Booth, B. M., & D'Mello, S. (2022). A Conceptual Framework for Investigating and Mitigating Machine-Learning Measurement Bias (MLMB) in Psychological Assessment. *Advances in Methods and Practices in Psychological Science*, *5*(1), 25152459211061337.

UNIVERSITÄT
DES
SAARLANDES

Danke für Ihre Aufmerksamkeit

Markus.langer@uni-saarland.de

Twitter: @LangersMarkus

UNIVERSITÄT
DES
SAARLANDES

Backup

Die Perspektive der Betroffenen – Studie I

- Beispiel Gerechtigkeitswahrnehmungen (Colquitt, 2001)
- **Informationelle Gerechtigkeit:** Es sollten ausreichende und vollständige Informationen vorliegen, die Entscheidungen und Ergebnisse erklären und rechtfertigen
- **Interpersonelle Gerechtigkeit:** Behandlung von Personen mit Würde und Respekt von Seiten derer, die Entscheidungen treffen und kommunizieren
- **Prozedurale Gerechtigkeit:** Wahrgenommene Gerechtigkeit des Entscheidungsprozesses
- **Distributive Gerechtigkeit:** Wahrgenommene Gerechtigkeit der Entscheidung an sich

hiips://www.jagreporter.af.mil/Post/Article-View-Post/Article/2544349/autonomous-weapons-need-autonomous-lawyers/

Beispiele Gerechtigkeit

- Geringe **Informationelle Gerechtigkeit**: „Sie werden von uns hören“
- Geringe **Interpersonelle Gerechtigkeit**: „Das Outfit steht Ihnen aber nicht wirklich“
- Geringe **Prozedurale Gerechtigkeit**: „Wir werden Ihnen viel schwerere Fragen stellen als den anderen Kandidaten“
- Geringe **Distributive Gerechtigkeit**: „Sie waren der beste Kandidat – aber wir haben uns für jemand anderen entschieden“