Unzulässige Diskriminierung durch KI-Systeme:

Technische Ursachen und Lösungsansätze

Bettina Berendt

TU Berlin, Weizenbaum Institut für die Vernetzte Gesellschaft, und KU Leuven

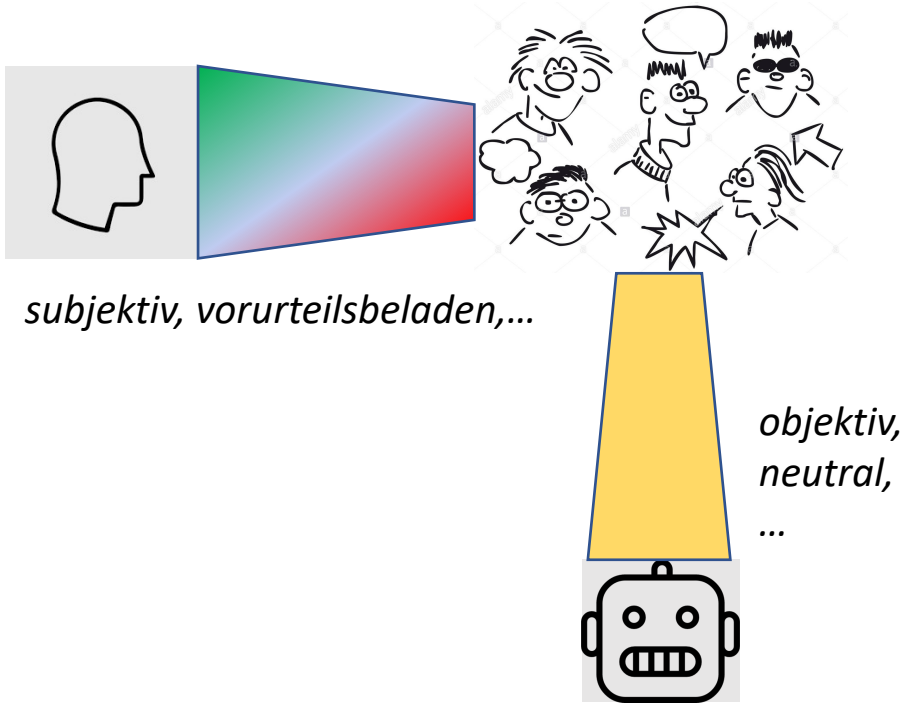
Folien und Literatur : www.berendt.de/bettina

Verbraucherrechtstage 2021

12.7.2021

Diskriminierung und die Hoffnung auf ihre Vermeidung durch Informatik / KI / ADMS

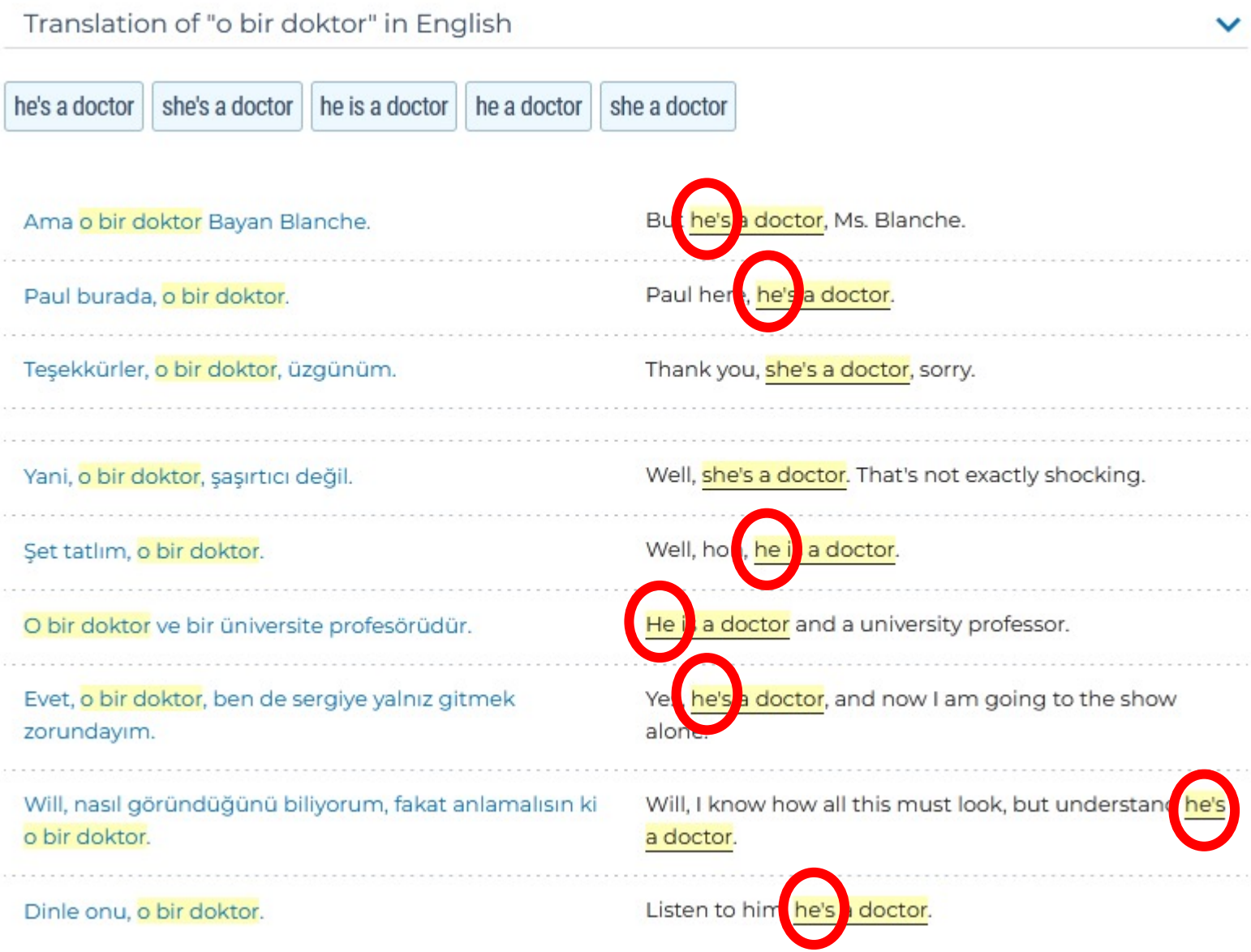
- Benachteiligungen aus Gründen der Rasse oder wegen der ethnischen Herkunft, des Geschlechts, der Religion oder Weltanschauung, einer Behinderung, des Alters oder der sexuellen Identität (§1 AGG)
- mittelbare Benachteiligung [:] dem Anschein nach neutrale Vorschriften, Kriterien oder Verfahren Personen [...] benachteiligen können, es sei denn, [...] rechtmäßiges Ziel sachlich gerechtfertigt und die Mittel [...] angemessen und erforderlich. (§3 (2) AGG)
- 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)
- "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."
  - Friedman and Nissenbaum (1996) Bias in Computer Systems



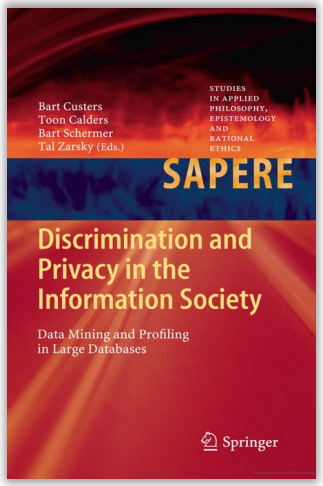
Bsp. 1: KI-Sprachverarbeitung
Bias in der maschinellen Übersetzung



Warum? Maschinelles Lernen aus Daten, die aktuelle & historische Ungleichheiten widerspiegeln



Gegenmaßnahmen: Änderungen an Daten und Algorithmen; Interdisziplinarität



FAIRNESS AND MACHINE LEARNING

Limitations and Opportunities

Solon Barocas, Moritz Hardt, Arvind Narayanan

Full book as PDF

The New York Times | https://www.nytimes.com/2021

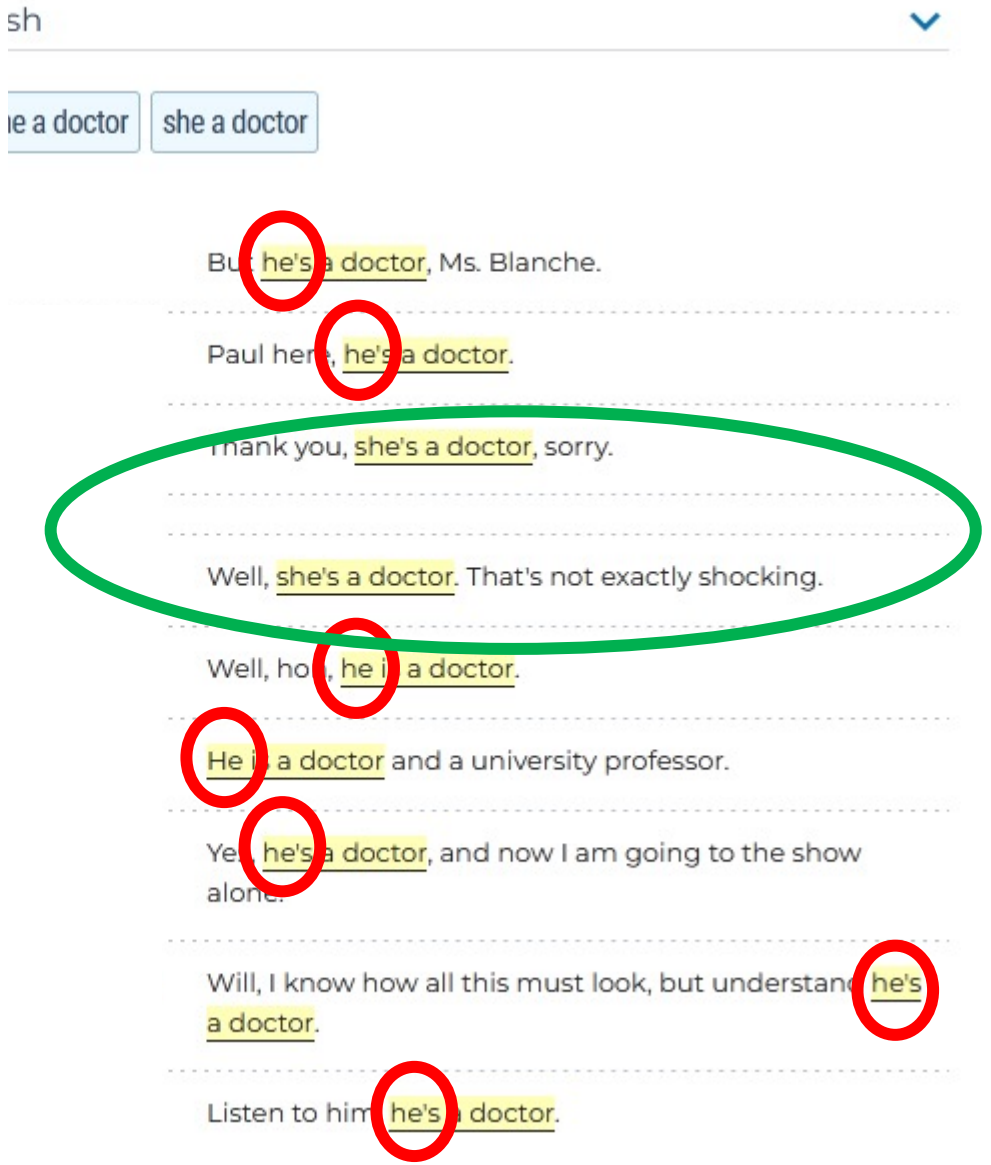
Using A.I. to Find Bias in A.I.

The problem of bias in artificial intelligence is facing increasing scrutiny from regulators and is a growing business for start-ups and tech stalwarts.

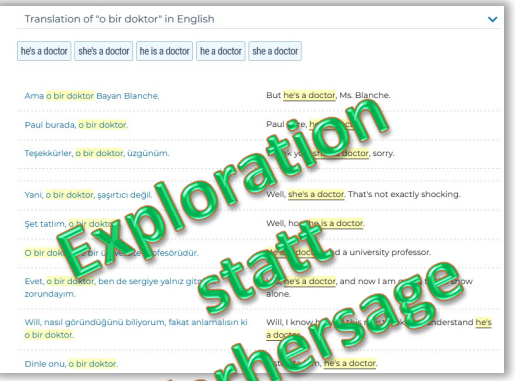


By Cade Metz

& & & ...



Einfluss von Algorithmus & Problemformulierung – Gegenmaßnahmen – Grenzen









Warum ist das ein Problem?

- “repräsentationaler Schaden”
  - Vorurteile in der Sprache & ihre Effekte
- “allokationaler Schaden”
  - Abgeleitete Entscheidungen
  - Z.B. Vorhersage von Interesse & “optimierte” Anzeigenschaltung
    - Vgl. Vortrag M. Spielkamp
  - Z.B. Vorhersage von Fähigkeiten & (Nicht-)Weiterbetrachtung von Bewerbungen
    - Achtung: Das Problem ist häufig nicht so direkt zu sehen wie in den hier gezeigten Beispielen! Typischer ist mittelbare Diskriminierung über Proxies (z.B. subtile “typisch weibliche Sprache”), die KI-Systeme lernen und die sehr schwer zu detektieren sind
  - Z.B. Vorhersage von Jobchancen & davon abhängiger Zugang zu Fördermaßnahmen
    - Achtung: Diese Systeme operieren bisher m.W. *nicht* auf Sprachdaten, aber grundsätzliche Probleme des maschinellen Lernens sind ähnlich

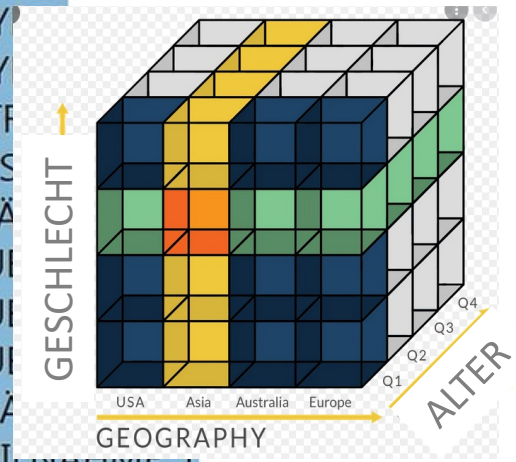
Bsp. 2: Der AMS-Algorithmus

- Projekt des Österreichischen Arbeitsmarktservices (AMS)
- Das „Arbeitsmarktchancen-Assistenz-System“soll auf Basis von Statistiken vergangener Jahre zukünftige Chancen von Arbeitssuchenden vorhersagen.
- Arbeitssuchende werden anhand ihrer „Integrationschance“ in drei Gruppen eingeteilt.
- Gefördert werden soll vorrangig im „mittleren“ Segment.
- Soziotechnische Analyse: Allhutter et al. (2020) – Schlussfolgerungen:
  - “Der IC-Wert hat weitreichende Konsequenzen in der AMS-Beratungspraxis und für Arbeitssuchende.
  - Transparenz und Einspruchsrechte sowie öffentliche Mitsprache [werden daher] gefordert.”



- Aktueller Status: AMS – Datenschutzbehörde – beim Bundesverwaltungsgericht

BE_INT
1 = f (0,10
2 – 0,14 x GESCHLECHT_WEIBLICH
– 0,13 x ALTERSGRUPPE_30_49
– 0,70 x ALTERSGRUPPE_50_PLUS
+ 0,16 x STAATENGRUPPE_EU
– 0,05 x STAATENGRUPPE_DRITT
3 + 0,28 x AUSBILDUNG_LEHRE
+ 0,01 x AUSBILDUNG_MATURA_PLUS
4 – 0,15 x BETREUUNGSPFLICHTIG
– 0,34 x RGS_TYP_2
– 0,18 x RGS_TYP_3
5 – 0,83 x RGS_TY
– 0,82 x RGS_TY
6 – 0,67 x BEEINTF
+ 0,17 x BERUFS
– 0,74 x BESCHÄ
+ 0,65 x FREQUE
+ 1,19 x FREQUE
+ 1,98 x FREQUE
7 – 0,80 x GESCHÄ
– 0,57 x MN_TEILNAHME_1
– 0,21 x MN_TEILNAHME_2
– 0,43 x MN_TEILNAHME_3)

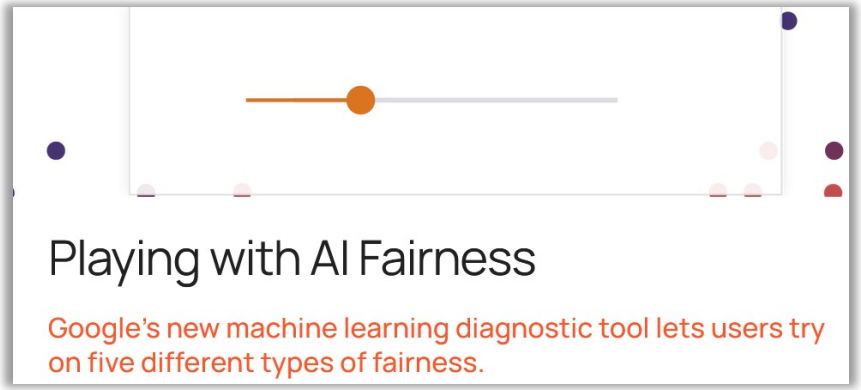


Wie könnte ein solches System “nicht-diskriminierend” sein?

- Sollte es für (z.B.) Männer und Frauen gleich oft das mittlere Segment vorhersagen?
- Sollten die Fehlerraten für diese Gruppen gleich sein?
- Sollte der Vergleich auf der Gruppen-Ebene oder auf der Ebene von ähnlichen Individuen stattfinden?

➔ Debatte um die Metriken für “Nichtdiskriminierung” bzw. “Fairness”

- Optimierungen für verschiedene Metriken stehen z.T. im Widerspruch
- Entspricht auch verschiedenen Formen von “Gerechtigkeit”
  - eine historisch sehr alte – und politische! – Debatte



https://pair-code.github.io/what-if-tool/ai-fairness.html

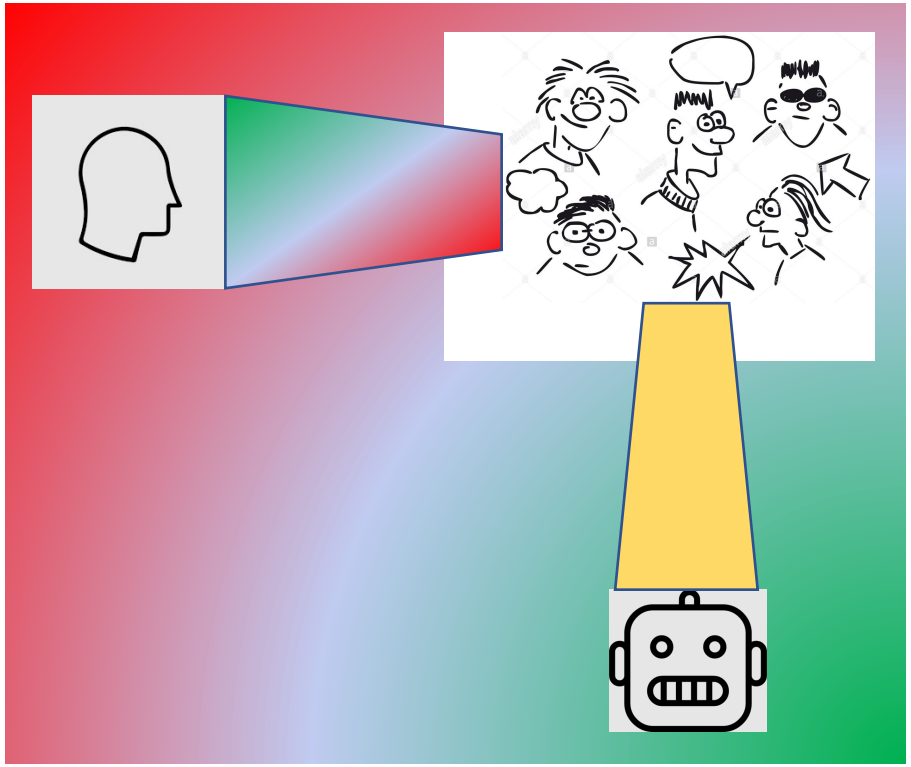
Gegenmaßnahmen: Was können KI-Forscher*innen beitragen? *(was sie auch tun)*

- Formalisierung von Metriken und Untersuchung ihrer formalen Eigenschaften kann Klarheit über Ziele erhöhen.
- Kritik der Annahmen in den Daten und der statistischen Eigenschaften des Systems → Einschätzung der Zuverlässigkeit der Vorhersagen.
- Wissen über den Unterschied zwischen Deskription und Vorhersage kann helfen, Aussagen wie “das reflektiert nur die harten Realitäten des Arbeitsmarktes” als zu kurz gegriffen zu kritisieren.
- Kritik der Interpretation: Es fehlen Daten über die Effektivität/Effizienz von AMS-Maßnahmen (und damit genau über die Intervention)!

Zooming out: Wo kommt die Diskriminierung her – und geht es nicht um mehr?

- Systeme wie das AMS basieren auf vielen Annahmen, z.B.
  - Risikofaktoren für Arbeitslosigkeit liegen (fast) nur beim Arbeitssuchenden
  - Die Kategorisierung von Bürgern nach ihrem ökonomischen Potenzial ist politisch vertretbar und sinnvoll/zielführend.
- Die Analyse und Reparatur ("De-Biasing") solcher Systeme macht weitere Annahmen, z.B.
  - Risikofaktoren für Bias / Diskriminierung liegen nur beim "Entscheider" (ob Mensch oder ADMS)
  - Strukturelle Diskriminierung wird dadurch vernachlässigt.
- Wer bestimmt das "Risiko" dieser Systeme?
- Wollen wir solche Systeme als Gesellschaft?

  (Scott et al., submitted; Delobelle et al., submitted: AU/BE/PL/DE)



Nein, ich habe kein Patentrezept zur Arbeitsmarktpolitik, aber: KI kann auch ganz anders genutzt werden!

Why is Swedish the top language in Sweden?



Duolingo's Swedish course turns out to be the most popular in Sweden itself: 27% of all users in Sweden are learning Swedish.

Why would that be the case? Immigration to Sweden has been skyrocketing in recent years: one in six Swedish residents in 2015 was born outside of Sweden. The fastest growing foreign-born groups are from Syria and Afghanistan, reflecting a recent increase in the refugee population. Duolingo recently released a Swedish course for Arabic speakers, which will hopefully help!

UNHCR and Duolingo Aim to Help More Refugees Enroll in Higher Education

Riham Alkousaa / 21 Jun 2021

https://blog.duolingo.com/which-countries-study-which-languages-and-what-can-we-learn-from-it/
https://www.al-fanarmedia.org/2021/06/unhcr-and-duolingo-aim-to-help-more-refugees-enroll-in-higher-education/

Fazit

- KI-Algorithmen können existierende Diskriminierungen replizieren und verstärken, und sie können neue schaffen.
- KI-Algorithmen können solche algorithmische Diskriminierung messen, verringern und manchmal ausschalten.
- Aber “Diskriminierung durch KI” findet im Kontext viel weiter reichender Strukturen von Ungleichheit und Diskriminierung statt.
- Die historische Chance dieser Situation ist groß
  - *Viele* Menschen (z.B. Informatiker*innen, oft in interdisziplinären Kontexten) denken aktiv über das Problem nach und handeln.
  - Wir haben als Gesellschaft und mit KI die Chance, Technologie anders als “konservativ” zu gestalten.

Literatur

Überblicksdarstellung, auch hinsichtlich der Beziehungen zum Recht, mit ausführlicher Bibliographie:

Berendt, B. (in press). Algorithmic discrimination. To appear in G. Comandè (Ed.), *Encyclopedia of Law and Data Science.* Edward Elgar Publishing Ltd. (last author version: https://people.cs.kuleuven.be/~bettina.berendt/Papers/Algorithmic Discrimination BB 2020 10 13.pdf)

In diesen Folien zitierte Literatur

Doris Allhutter, Astrid Mager, Florian Cech, Fabian Fischer, and Gabriel Grill. 2020. Der AMS Algorithmus - Eine Soziotechnische Analyse des Arbeitsmarktchancen-Assistenz-Systems (AMAS). Technical Report. Wien: ÖAW, ITA 2020-02. http://epub.oeaw.ac.at/?arp=0x003bfdf3/

Solon Barocas, Moritz Hardt and Arvind Narayanan (2019). Fairness and Machine Learning. http://www.fairmlbook.org

Custers, B., Calders, T., Schermer, B., Zarsky, T. (Eds.) (2013). Discrimination and Privacy in the Information Society. Data Mining and Profiling in Large Databases. Springer

Pieter Delobelle, Kristen M. Scott, Sonja Mei Wang, Milagros Miceli, David Hartmann, Tianling Yang, Elena Murasso, Karolina Sztandar-Sztanderska and Bettina Berendt (submitted). Time to Question if We Should: Data-Driven and Algorithmic Tools in Public Employment Services. *Available from the authors per email request.*

Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. ACM Transactions on Information Systems 14, 3 (July 1996), 330–347. https://doi.org/10.1145/230538.230561

Kristen M. Scott, Pieter Delobelle, Sonja Mei Wang, Milagros Miceli, David Hartmann, Tianling Yang, Elena Murasso, Karolina Sztandar-Sztanderska and Bettina Berendt (submitted). Data-Driven and Algorithmic Tools in Public Employment Services: Towards a Stakeholder-Centric Perspective. *Available from the authors per email request.*

Zusätzliche Folien

EU: “Hochriskante” diskriminierende KI-Systeme: Verbot oder “Reparatur”? Geht das?

- (17) KI-Systeme, die **von Behörden oder in deren Auftrag das soziale Verhalten natürlicher Personen für allgemeine Zwecke bewerten**, können zu diskriminierenden Ergebnissen und zur Ausgrenzung bestimmter Gruppen führen. Sie können die Menschenwürde und das Recht auf Nichtdiskriminierung sowie die Werte der Gleichheit und Gerechtigkeit verletzen. [...] kann zu einer [...] Benachteiligung bestimmter natürlicher Personen oder ganzer Gruppen natürlicher Personen [...] führen, die im Hinblick auf ihr soziales Verhalten oder dessen Tragweite ungerechtfertigt oder unverhältnismäßig ist. Solche KI-Systeme sollten daher **verboten** werden.
- (35) KI-Systeme, die in der allgemeinen oder beruflichen **Bildung** eingesetzt werden, insbesondere um den Zugang von Personen zu Bildungs- und Berufsbildungseinrichtungen oder ihrer Zuordnung dazu zu bestimmen oder um Personen im Rahmen von Prüfungen als Teil ihrer Ausbildung oder als Voraussetzung dafür zu bewerten, sollten als **hochriskant** angesehen werden. [Sie können] das Recht auf Nichtdiskriminierung verletzen und historische Diskriminierungsmuster fortschreiben.
- (36) KI-Systeme, die in den Bereichen **Beschäftigung, Personalmanagement und Zugang zur Selbstständigkeit** eingesetzt werden [...] sollten ebenfalls als **hochriskant** eingestuft werden, [...] historische Diskriminierungsmuster fortschreiben.

Tackling gender, race and other biases in AI (RIA)

TOPIC ID: HORIZON-CL4-2021-HUMAN-01-24

ExpectedOutcome:

Proposal results are expected to contribute to the following expected outcomes:

- Increased availability and deployment of unbiased and bias-preventing AI solutions across a wide range of industrial and digital sectors
- AI-based solutions for enhancing digital equality and social inclusion for women and girls, and other groups at risk of discrimination, such as ethnic minorities and the LGBTIQ community
- Increased involvement of underrepresented persons in the design, development, training and deployment of AI.
- Increased awareness, knowledge and skills about trustworthy, bias-free and socially responsible AI in the tech industry and scientific community

Vorschlag für eine VERORDNUNG DES EUROPÄISCHEN PARLAMENTS UND DES RATES ZUR FESTLEGUNG HARMONISIERTER VORSCHRIFTEN FÜR KÜNSTLICHE INTELLIGENZ (GESETZ ÜBER KÜNSTLICHE INTELLIGENZ) UND ZUR ÄNDERUNG BESTIMMTER RECHTSAKTE DER UNION , COM(2021) 206 final

Zooming out (1): Ist es wirklich nur die KI?



Betreuungspflichten werden nur für Frauen abgefragt!