DISKRIMINIERENDE ALGORITHMEN: TECHNISCHE ODER SOZIALE FRAGE?

17. Juni 2021

Linus Petermann & David Sommer

Automatisierte Selektion in Bewerbungsverfahren

- Einfaches Beispiel:
  - *Algorithmus: Vorselektion Bewerbungsgespräche*
  - *Trainiert an alten Bewerbungsunterlagen, sucht Muster für «eingestellt»*
  - *Optimales Muster «eingestellt»: Schweizer Pass, Wohnort Zürich, Zweisprachig*
- Aktuelle Bewerber*innen:

Nachname	*Nationalität*	*Letzter Bildungsabschluss*	*Sprachen*	*Wohnort*	*#calc Distanz zu Arbeitsort (km)*
Pasquinelli	Schweiz	MA Universität Zürich	Deutsch, Englisch	Zürich	2
Uljanow	Schweiz, Russland	MA ZHAW	Deutsch, Englisch, Russisch	Spreitenbach	10
Lopez	Spanien	MA Universität Madrid	Deutsch, Englisch, Spanisch	Spreitenbach	10

Direkte und indirekte Diskriminierung

- Selektion nach Nationalität → Direkte Diskriminierung
  - *Lösung: Attribut entfernen*
- Selektion nach Wohnort → Indirekte Diskriminierung
  - *Starker Zusammenhang mit nicht-schweizerischer Nationalität*
  - *Lösung?*

→ Reproduktion sozialer Normen

Nachname	*Nationalität*	*Letzter Bildungsabschluss*	*Sprachen*	*Wohnort*	*#calc Distanz zu Arbeitsort (km)*
Pasquinelli	Schweiz	MA Universität Zürich	Deutsch, Englisch	Zürich	2
Uljanow	Schweiz, Russland	MA ZHAW	Deutsch, Englisch, Russisch	Spreitenbach	10
Lopez	Spanien	MA Universität Madrid	Deutsch, Englisch, Spanisch	Spreitenbach	10

Wandelnde Normen, unkritische Algorithmen

- Früher: Ungleichbehandlung von Frauen in jungem Alter
  - *Gerechtfertigt und mit «sachlichem Grund»*
- Heute: Rechtfertigung wird zunehmend prekär

Nachname	*Nationalität*	*Letzter Bildungsabschluss*	*Sprachen*	*Geschlecht*	*Alter*
Pasquinelli	Schweiz	MA Universität Zürich	Deutsch, Englisch	w	30
Uljanow	Schweiz, Russland	MA ZHAW (Fachhochschule)	Deutsch, Englisch, Russisch	m	32
Lopez	Spanien	MA Universität Madrid	Deutsch, Englisch, Spanisch	w	43

Soziale Prägung der Technik

- Diskriminierung nicht «nur» technisches Problem
- Bestimmung von Diskriminierung ist offen und wird gesellschaftlich ausgehandelt
  - *Was ist «sachlich»?*
- Vorstellungen und Ordnung von Welt prägen Algorithmen
  - *Algorithmus richtet sich nach vorhandenen Zusammenhängen*

  → *Immer eine Übertragung von Werten und Normen in Technik*

Aufbau

- ADM Systeme: Begriff, Einsatz, Verbreitung
- Technische Grundlagen: Funktionsweise, Problematiken und Widersprüche
- Datenschutz und seine Grenzen
- Rechtliche Grundlagen und mögliche Rahmenbedingungen
- Was tun? Perspektiven für Gesellschaft und Politik
- Fazit

ADM Systeme: Begriff, Einsatz

- KI, AI, Machine Learning, Neuronale Netzwerke…
  - *Unklare / verschwommene Definitionen*

→ Automated Decision Making (ADM Systeme)

- Heutiger Haupteinsatzzweck
  - *Effizienzsteigerung in Prozessen*
    - Ermöglicht Verarbeitung viel grösserer Datenmengen
  - *"Faire" Beurteilung*
    - Anstatt viele varierende Expertenbeurteilungen

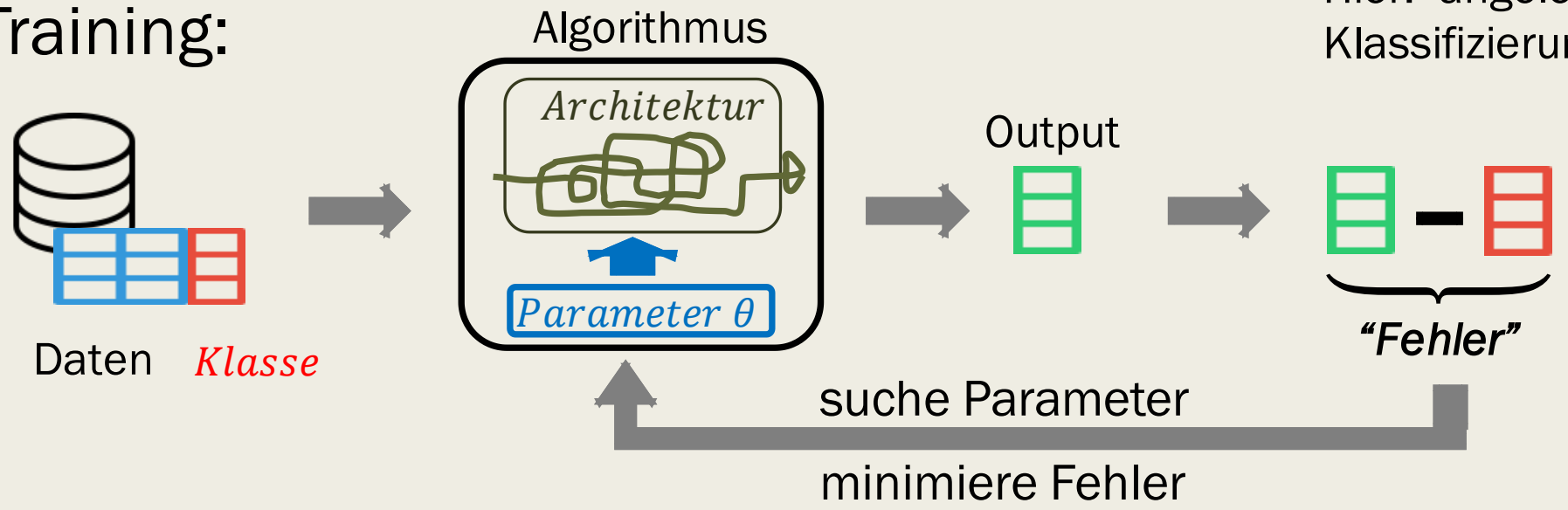
ADM Systeme: Verbreitung

- Grosse Bandbreite
  - *Automatisierung von Entscheidungs- und Dienstleistungsprozessen*
- Unterschiedliche Folgen und Diskriminierungseffekte
- Beispiele:
  - *Amazon Bewerbungsverfahren*
    - Ausschluss nicht-männlicher Personen (2018)
  - *Kindsgefährdungsanalysen New York*
    - Unterstützung von Präventiveingriffen
  - Zuteilung von Schulhäusern
    - Höhere soziale Heterogenität
    - Z.B. Versuchsprojekt in Zürich

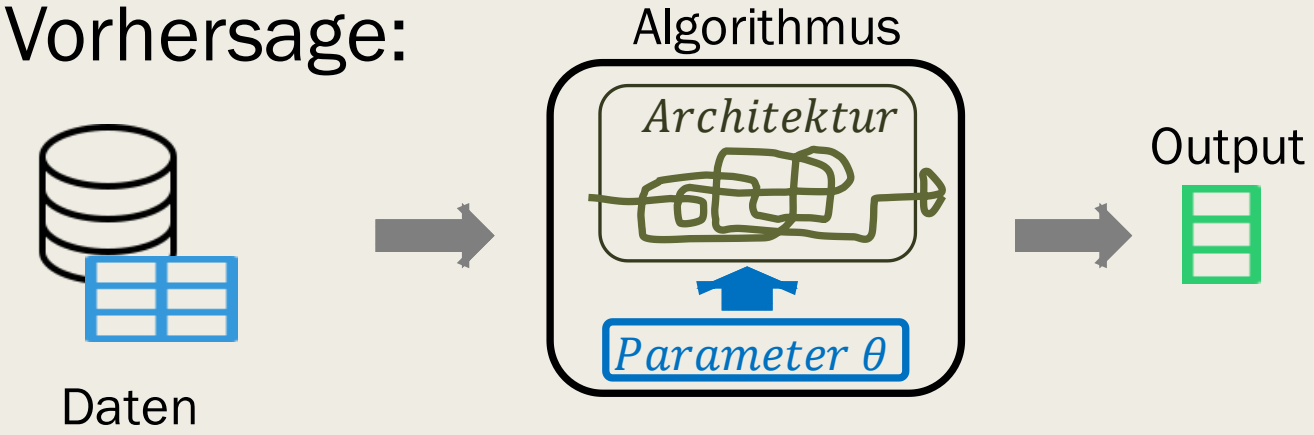
Wie funktionieren diese Algorithmen?

Training:

Hier: angeleitete (supervised) Klassifizierung



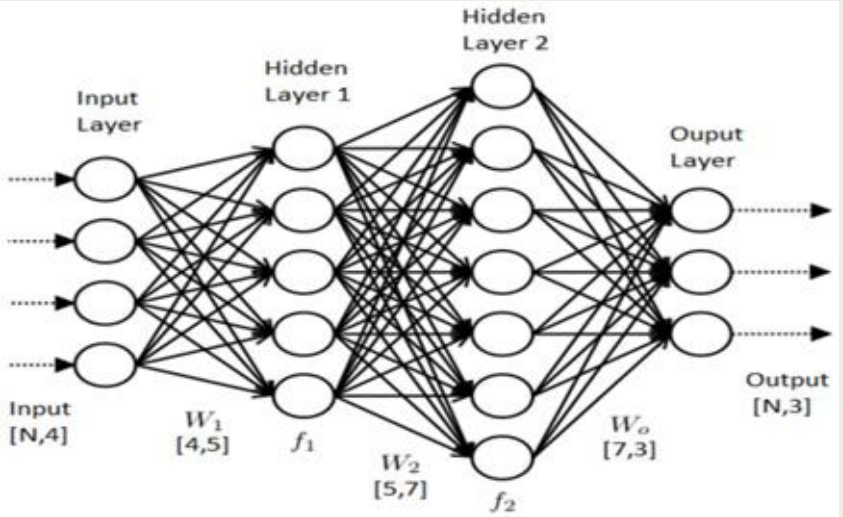
Vorhersage:



Qualitätsmessung, oft: Vorhersagegenauigkeit (im Schnitt)

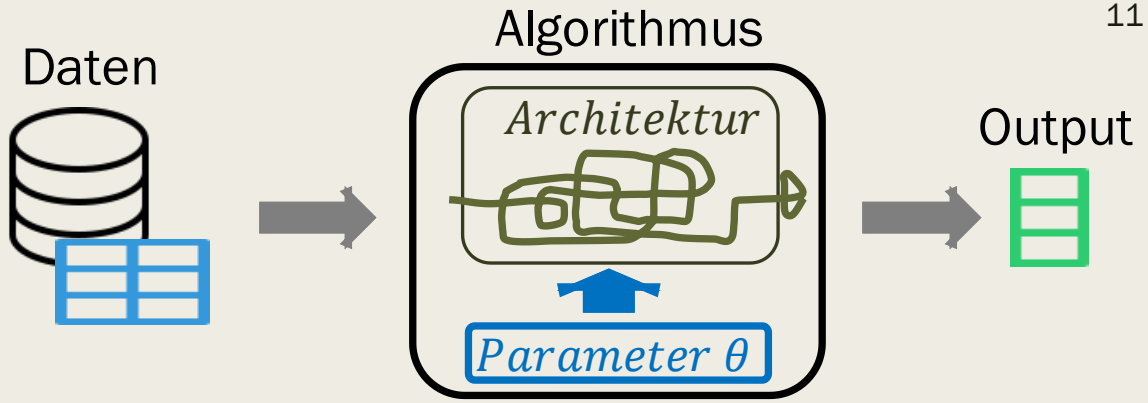
Kausalität

- Kausalität:
  → *Warum ein Algorithmus wie klassifiziert*
- Sinn ist nicht immer ersichtlich
  → *Algorithmus findet unbekannte Korrelation*
- Kann auch falsche Schlüsse ziehen
  → *Wolf wird an Schnee im Hintergrund erkannt*
- Zunahme Komplexität: Z.B. Neuronale Netze:
  → *Milliarden an Parameter*
  → *Kausalität schwierig*



Quelle Bilder:
https://s3.amazonaws.com/media.eremedia.com/wp-content/uploads/sites/4/2019/12/16105353/AI-wolf-or-husky.jpg
https://miro.medium.com/max/700/1*ZB6H4HuF58VcMOWbdpcRxQ.png

Bias (Verzerrung)



- Rohdaten beinhaltet bereits Diskriminierung
  - *Früher wurden mehr Schweizer eingestellt*
    ⟶ Algorithmus empfiehlt eher Schweizer
- *Population Bias*: Untergruppen waren nicht Teil des Trainings:
  - *Seifenspender funktioniert bei dunkler Hautfarbe nicht*
- *Representation Bias*: Untergruppen werden vernachlässigt
  - *Niedriger Anteil ⟶ grosser "Fehler" Einzelner im Schnitt vernachlässigbar*
- *Algorithmic Bias*: Architektur-/Design-Entscheidungen
  - *Algorithmus schafft es nicht, auf wichtige Details einzugehen*
  - *Definition von Fehler hat enorme Auswirkungen auf Resultat*

Fairness?

- Angesprochene Algorithmen unterscheiden zwingend
- Wie bekommt man diese “fair”
  - *Anpassung der Rohdaten*
  - *Korrektur beim „Fehler“*
- Was ist fair?
- Individuelle vs. Gruppen-Fairness:
  - *Bsp: Priorisierte Einstellung von Frauen in Männerdomäne*
  - *Für Parität im Schnitt: Frau hat deutlich höhere individuelle Einstellungs-Chancen*
- Generell: Prioritisierung eines Zusammenhangs schwächt meistens einen anderen
  - Grund: andere Daten-Ausgangsstruktur --> andere Dynamik
- Trade-off zwischen Effizienz und Fairness
  - *Schwächt Nützlichkeit des Algorithmus*
- Algorithmen zwar neutral, reproduzieren aber Datenstruktur und "künstliche" Zwänge

→ Fairness ist technisch nicht eindeutig, benötigt (gesellschaftliche) Diskussion

Rechtliche Grundlagen und mögliche Rahmenbedingungen

Datenschutz!

- Idee: Schutz des Rechts auf informationelle Selbstbestimmung & Wahrung der Privatsphäre
- Praxis: Schutz vor missbräuchlicher Datenverarbeitung & Datenauskunftsrecht
  - *Datensammlung und –nutzung = Machtfrage*
    - Überwachung, z.B., Fichen-Affären (CH)
    - Missbrauch und Manipulation, z.B. Cambridge Analytica
- Datenschutzgesetze verlangen oft
  - *Zweckbindung, Korrektheit, Rechenschaftspflicht, zeitliche Speicherbegrenzung*

Datenschutzgesetze und ADM

- **Schweiz:** Neues Datenschutzgesetz, Art. 21 (ab 2022)
  - *Recht auf Informierung und Überprüfung bei / von vollautomatisierten Entscheiden*
- **Europäische Union:** DS-GVO Art. 22 Abs. 1
  - *Recht auf nicht ausschließlich automatisierte Verarbeitung bei rechtlicher Wirkung*

⟶ Enge Definitionen, schränken nur Extremfälle ein, nur Einzelfallperspektive

- Informationen / Daten sind an verschiedenen Orten gespeichert
  - *Schwierigkeit in der Übersicht über effektive Datenverwendung und -speicherung*
  - *Daten werden dort auch datenschutzkonform genutzt / von Algorithmen verwertet*
- Datenschutz greift nur bei persönlichen Daten
  - *Bsp. nicht bei: Optimale Platzierung von Feuerwehrwachen (basierend auf öffentlichen Daten)*

Hilft Datenschutz bei AMD Systemen?

- Algorithmen können trainiert werden, ohne Privatsphäre Einzelner zu verletzen (z.B. Differential Privacy) → Keine Verhinderung von Gruppenbenachteiligung
- Probleme des individualistischen Datenschutzkonzepts
  - *Datenschutz heute auf individuelle Personenfälle fokussiert, aber Effekte häufig auf Gruppenebene*
  - *+ Einforderung von Rechten setzt Kompetenzen und Ressourcen voraus → Rechte sollte nicht nur für die die gelten, die sich wehren können*
- → Hin zu kollektivem Rechtsverständnis und -schutz (z.B. Sammelklagen)

→ **Kein ausreichender Schutz durch Datenschutz!**

  - *Datenschutz verminderte extensives Verteilung von (potentiell fehlerhaften) Informationen*
  - *Datenschutz ist ein Schritt in die richtige Richtung, reicht aber nicht.*

Prinzipien der ADM-Regulation

- Transparenz beim Einsatz
  - *Kennzeichnung, komplette Veröffentlichung?*
- Recht der Zivilgesellschaft auf Einblick und nach Rechtfertigung durch Verantwortliche
- **Sanktionen!**
  - *Klare Haftung, von Menschen oder Unternehmen*
  - *Keine Auslagerung an andere Juristriktionen (Google vs. Spanien,* CJEU, 2014*)*

Stand ADM Regulationen

- **EU**: Gesetzesvorschlag veröffentlicht 21. April 2021 [0]
  - *Biometrische Massenüberwachung unzureichend verboten („Reclaim your face „)*
- **Schweiz**: diverse Strategiepapiere [1,2], kein konkreter Gesetzesvorschlag
  - *Europarat (Nicht EU): Committee on Artificial Intelligence (CAHAI) [3]*
  - *Schweiz ist Mitglied (Exekutive ist involviert)*
  - *29. April 2021: Multi-Stakeholder Consultation geschlossen*

[0] EU-Rat: Proposal for a Regulation laying down harmonised rules on artificial intelligence (2021)
[1] Bericht an Bundesrat: Herausforderungen der künstlichen Intelligenz (Ende 2019)
[2] Leitlinien für den Umgang mit künstlicher Intelligenz in der Bundesverwaltung (November 2020)
[3] https://www.coe.int/en/web/artificial-intelligence/cahai

Stand ADM Regulationen II

- Generell: 3-4 Kategorien
  - *Low risk: keine/wenig Regulation (z.B. Korrektheitskontrollen von Formularen)*
  - *High risk: weitgehende Regulationen (z.B. gezielte politische Werbung)*
  - *Verboten: z.B. Biometrische Identifikation für Massenüberwachung*

- Zwei Ansätze scheinen sich durchzusetzen:
  - *Kritikalitätsansatz: Regulierungsanforderungen je nach Schadensrisiko*
  - *Schadensansatz: Bei Verstoss kann geklagt werden (privatwirtschaftlich)*

- Unterscheidung zwischen Staat und Privatwirtschaft
  - *Staat arbeitet mit Verfügungen auf Rechtsgrundlage, Zwang*
  - *„Freier“ Vertrag, mehrere Anbieter*innen*
- Idee: Keine Unterscheidung zwischen Entscheiden von Mensch oder Maschine

Gesellschaftlicher Umgang mit ADM Systemen

- Weiterentwicklung und schnelle Verbreitung von ADM Systemen absehbar
- **Schleier der Neutralität und Technikgläubigkeit ablegen:**
  - *Verschleiert Risiken*
  - *Werte- und Normübertragungen in Technik sichtbar machen und reflektieren*
- Gefahr: Verstärkung sozialer Normen, versteckte Diskriminierung
  - *Algorithmen ohne «Gewissen» → Gefahr der Verstärkung sozialer Normen*
  - *Konventionsdruck*
- ADM Systeme zum Teil von Debatten machen:
  - *Datenbasis: Welche/Wessen Daten, welche Filter*
  - *Struktur/Architektur: Art des Algorithmus, Erneuerungszeitpunkte*
  - *Einsatz: Wann, Wofür, Wie? →* ***Ob überhaupt?***
- Positiver Seiteneffekt: verbesserte statistische Erkennbarkeit von Diskriminierung

Fazit

- Reproduktion und Verstärkung sozialer Normen und bestehender Ungleichbehandlung
  - *Diskriminierende Gesellschaft → diskriminierende Daten & Algorithmen*
- Technische Lösungen beschränkt
  - *Diskriminierung technisch nicht komplett auflösbar*
  - *Widersprüche in Fairnessdefinitionen*
  - *→ Gesellschaftliche „Begleitung“ von und Entscheidung über technischen Entwicklungen notwendig*
- Technikgläubigkeit verführt zur Verantwortungsabgabe
- Regulierungsrahmen ungenügend
  - *Intensive Diskussionen, Regulierungen absehbar*
  - *Zivilgesellschaft vs. Unternehmens- und Regierungsbestrebungen stark asymmetrisch*
  - *Datenschutzansätze nicht ausreichend*
  - *Hin zu kollektiven Rechtsverständnis (kollektiver Rechtsschutz, Sammelklagen)*

Weiterführende Links und Literatur

Hagendorff, Thilo (2019): *Maschinelles Lernen und Diskriminierung: Probleme und Lösungsansätze*, in: *Österreichische Zeitschrift für Soziologie*, Jg. 44, Nr. 1, S. 53–66, doi: 10.1007/s11614-019-00347-2.

Orwat, Carsten (2019): *Diskriminierungsrisiken durch Verwendung von Algorithmen*, (Studie) Berlin: Antidiskriminierungsstelle des Bundes.

Automating Society Report Schweiz (2020) - AlgorithmWatch /CH

Sommer, David (2021): *Machine-Learning-Leitlinien des Bundes*, [online] https://www.digitale-gesellschaft.ch/2020/12/17/machine-learning-leitlinien-des-bundes-entscheidende-gefahren-nicht-benannt/

Tisne, Martin (2021): *Kollektiver Datenschutz: Was dahinter steckt und warum er nötig ist*, Heise online, (https://www.heise.de/hintergrund/Kollektiver-Datenschutz-Was-dahinter-steckt-und-warum-er-noetig-ist-6054822.html)

Example representation bias

Quelle: Mehrabi, Ninareh, et al. "A survey on bias and fairness in machine learning." *arXiv preprint arXiv:1908.09635* (2019).