

Entstehung und Veränderung von Diskriminierungsrisiken durch KI

Kurzvortrag, 83. Deutscher Fürsorgetag, 17. Sep. 2025, Erfurt

Dr. Carsten Orwat

Institut für Technikfolgenabschätzung und Systemanalyse (ITAS)
des Karlsruher Instituts für Technologie (KIT)



Algorithmische Entscheidungen

Entscheidungsregeln in halb- oder vollautomatisierten Entscheidungen:

- In regel-basierter KI (z.B. Expertensystemen) „per Hand“ programmiert
 - Prinzipiell könnte man Kenntnis von Parametern und Regeln haben
- Mit maschinellem Lernverfahren entwickelt (Ermittlung von Verhältnissen zwischen Parametern und Korrelationen in Trainingsdatensätzen):
 - Auf Basis vorhergehender Entscheidungen (z.B. alle Bewerbungen von später „erfolgreichen“ Arbeitnehmer*innen)
 - An Hand von (neuen) Messgrößen für soziale Konstrukte (z.B. zur Abbildung von Persönlichkeitseigenschaften „Gewissenhaftigkeit“ und „Nicht-Neurotizismus“ als Hinweise auf Arbeitsleistung¹)
 - Komplexität durch sehr große Anzahl von Parametern; führt oft zur Intransparenz von Parametern und Regeln
- Generative KI zunehmend auch bei Entscheidung über Menschen (z.B. bei Diagnosen und Therapiewahl)²

¹ Aus Köchling et al. (2021) (Replikationsexperiment zu Systemen der Human Resource Analytics (HireVue, Precire etc.)).

² Siehe Moulaei et al. (2024), Wang et al. (2025).

Typen von Bias bei maschinellem Lernen (Auswahl)¹

Bei Datenzusammenstellung:

- **Historischer Bias:** soziale Ungleichheiten oder Stereotypen in Datensätzen, z.B. stereotype Wortzusammenhänge bei Natural Language Processing
- **Repräsentationsbias:** Datensatz enthält nicht bestimmte Teile der Bevölkerung oder Population hat sich über die Zeit geändert, z.B. mangelnde ethnische Diversität in Bilddatensätzen
- **Bias bei Messgrößen:** ungeeignete Merkmale oder Label (zu messende Hilfsgröße), um das soziale Konstrukt (Differenzierungsziel) abzubilden
 - Hilfsgröße ist eine zu starke Vereinfachung für Differenzierungsziel, z.B. unmöglich „erfolgreiche Studierende“ in einer Variabel umzusetzen
 - Eignung der Methode oder Genauigkeit der Messung variiert zwischen Gruppen

¹ Auswahl basierend auf Suresh & Guttag (2021) und Mehrabi et al. (2021); Beispiele für algorithmische Bias oder Diskriminierungen in AIAAIC Repository oder Orwat (2019).

Typen von Bias bei maschinellem Lernen (Auswahl)

Bei Modellanwendung:

- **Bei Erlernen bzw. Auswahl von Modellen:** wenn Priorisierung eines Ziels ein anderes schädigt, z.B. Schutz der Privatsphäre kann Genauigkeit vermindern
- **Bias bei Evaluierung:** der Datensatz zur Überprüfung repräsentiert nicht die Population, für die das Model genutzt wird (z.B. Unterrepräsentation von dunkelhäutigen Frauen in Benchmark-Datensatz führt zu schlechterer Leistung von Gesichtserkennungssystemen bei ihnen)
- **Populationsbias:** Modell wird bei einer Population angewandt, die von der Population, an der Modell entwickelt wurde, abweicht

Fazit:

- Ursachen von Diskriminierungsrisiken mehr als „Datenbias“
- Diskriminierungsrisiken entstehen durch Entscheidungen von Entwickelnden und Anwendenden

Von Bias zur Diskriminierung

Diskriminierung nach rechtlicher Definition:

- Diskriminierung ist eine ungerechtfertigte Ungleichbehandlung mit Bezug auf ein geschütztes Merkmal (z.B. ethnische Herkunft, Geschlecht, Religion oder Weltanschauung, Behinderung, Alter, sexuelle Identität)¹
 - Differenzierung mit Verwendung des geschützten Merkmals (direkte Diskriminierung) oder
 - scheinbar neutrale Regeln benachteiligen doch entlang geschützter Merkmale (indirekte Diskriminierung)
- Behandlung muss Betroffene im Vergleich zu anderen schlechter stellen
- Gerechtfertigte Ungleichbehandlung in Ausnahmefällen²

¹ Nach § 1 Allgemeines Gleichbehandlungsgesetz (AGG).

² §§ 3(2), 5, 8-10, 19 und 20 AGG.

Systemische Diskriminierungsrisiken

Systemische Risiken: als potentielle Schäden auf gesellschaftlicher Ebene, die vor allem durch die Strukturen und Operationen eines gesellschaftlichen Systems verursacht werden

- Marktkonzentration bei KI-Anbietern und „algorithmischen Monokulturen“ mit wenigen oder ähnlichen KI-Modellen (vor allem der generativen KI):
 - Auch kleine Fehlerraten führen zu großen gesellschaftlichen Wirkungen
 - Schwere der Schäden wächst nicht-linear mit abnehmenden Ausweichmöglichkeiten für Betroffene¹
- Integration von Modellen der generativer KI in weitere IT-Systeme
 - Fehlerfortsetzung und Multiplikationseffekt
- Rückkopplungsschleifen durch „Wiederverwendung“ von Daten über Ergebnisse als weitere Trainingsdaten (z.B. bei predictive policing): können Bias vergrößern²

¹ Toups et al. (2024).

² Taori and Hashimoto (2023).

Systemische Diskriminierungsrisiken

- Ungeeignete Metriken bei der Entwicklung von KI-Modellen
 - Bilden starke Anreize für Ausrichtung der Entwicklung
- Schwächen des Antidiskriminierungs- und Datenschutzrechts (mittlerweile unzureichender Individualrechtsansatz) sowie der KI-Verordnung

Risikoverminderung und Restrisiken

Restrisiken von algorithmischen Bias

- Europäische KI-Verordnung:
 - Verweis auf Maße an Genauigkeit, Benchmarks und Metriken¹ (können Fairnessmetriken mit Fehlerraten enthalten)
 - Anforderungen an Risikomanagement: in dem Maße, so dass „Restrisiko“ als „vertretbar“ beurteilt werden kann² sowie „angemessenes Maß an Genauigkeit“³: Abwägungen zwischen Effizienz und Grundrechtseinschränkungen wird für Anbieter möglich
- Restrisiken sind wahrscheinlich:
 - Dadurch gesellschaftlich ungeklärte Niveaus von hinzunehmenden Diskriminierungsrisiken (es liegen keine allgemeinen Grenzwerte vor)
 - Mit Reichweiteneffekt und systemischen Risiken: auch scheinbar kleine Fehlerraten können zu Diskriminierung von großen Populationen führen

¹ Artikel 13(3)b ii und Artikel 15(2) KI-Verordnung

² Artikel 9(5) KI-VO.

³ Artikel 15(1) KI-VO.

Faktoren der Verletzung der Menschenwürde

Potentielle Verletzung der Menschenwürde bei algorithmischen Entscheidungen

- Behandlung von Personen als bloßes Objekt, Betroffene werden nicht als Individuum wahrgenommen, stattdessen meist Generalisierungen anhand von Daten über Gruppen

Kompensation bei Verletzung der Menschenwürde

- Moralische verwerfliche Behandlung wird üblicherweise mit informierter **Zustimmung** in moralisch akzeptable Behandlung transformiert
 - Voraussetzungen für valide Zustimmung: vor allem Freiwilligkeit und Wahlmöglichkeiten, Verständnis der Behandlung und deren Konsequenzen durch die Betroffenen bzw. Zustimmunggebenden

Faktoren der Verletzung der Menschenwürde

- **Aber bei algorithmischen Entscheidungen:** Eingeschränkte Möglichkeiten der Zustimmung durch Betroffene
 - Ohnehin Überlastung oder Versagen der „informierten Einwilligung“ im Datenschutzrecht¹
 - Teilweise Vorrang von Betriebs- und Geschäftsgeheimnissen
 - Mangelnde Nachvollziehbarkeit wegen intransparenter Entscheidungskriterien bei komplexen ML-Verfahren und generativer KI
 - Erklärbarkeit, Beeinflussbarkeit und Anfechtbarkeit fraglich
 - Wissenschaftliche Diskussion und Validierung der Entscheidungskriterien bleibt aus
 - Zunehmend Zwangslagen (abnehmende Ausweichmöglichkeiten, strukturelle Dominanz)

¹ Zusätzlich: Artikel 13 i.V.m. 22 DSGVO zum Recht auf Informationen zur involvierten Logik bei automatisierten Entscheidungen sowie Artikel 86 KI-VO zum Recht auf Erläuterung der Rolle und wichtigsten Elementen des KI-Systems bei Entscheidungen.

Faktoren der Verletzung der Menschenwürde

Weitere Verletzungsmöglichkeit:

- Schwerwiegende oder strukturelle Diskriminierungen ist eine Menschenwürdeverletzung:¹
 - Werden mit Zunahme systemischer Risiken wahrscheinlicher

¹ Siehe „NPD-Urteil“ BVerfGE 144, 20, Rn. 541; Höfling (2021) Rn. 35; Herdegen (2022) Rn. 120; Jarass (2022) Rn. 12; Dreier (2018) Rn. 138; Lehner (2013) S. 226-248; Langenfeld (2022) Rn. 98; u.a. Weitere Merkmale des Art. 3(3) GG sind Sprache, Heimat und Herkunft, Glaube und religiöse oder politische Anschauung.

Vielen Dank für Ihre Aufmerksamkeit!

Kontakt: orwat@kit.edu

Entscheidungen mit KI (vereinfachte Darstellung)

	Regelbasierte KI	Prädiktive KI	Generative KI
	Z.B. als Expertensysteme	Z.B. zur Risikoabschätzung im Marketing, Kreditvergabe, Personalwesen, Justizsystem	Zur Erzeugung von Inhalten (z.B. Chatbots), zunehmend bei Entscheidungen über Menschen
Zwecksetzungen	Meist für einen Zweck	Meist für einen Zweck	Auch als KI-Modelle mit allgemeinem Verwendungszweck, Einbau in weiteren IT-Systeme (z.B. zur Antragsprüfung)
Entscheidungsregeln:	Programmiert, Anzahl von Variablen eingeschränkt	Trainiert mit Erkennen von Mustern bzw. Korrelationen in Datensätzen bestimmter Domänen	Trainiert i.S.v. Bilden von wahrscheinlichen Sprachmustern in sehr großen Datensätzen oder anderen Modalitäten (Video, Audio)
Nachvollziehbarkeit, um Regeln den Betroffenen zu erklären:	Prinzipiell vollständig	Abnehmend mit zunehmender Komplexität; gegebenenfalls mit „Explainable AI“	Nein
Diskriminierungen erkennbar:	Prinzipiell im Code erkennbar	Mit zunehmender Komplexität nur mit Erfassung der Ergebnisse erkennbar	Nur mit Erfassung der Ergebnisse erkennbar, ggf. auf Angaben der Hersteller angewiesen