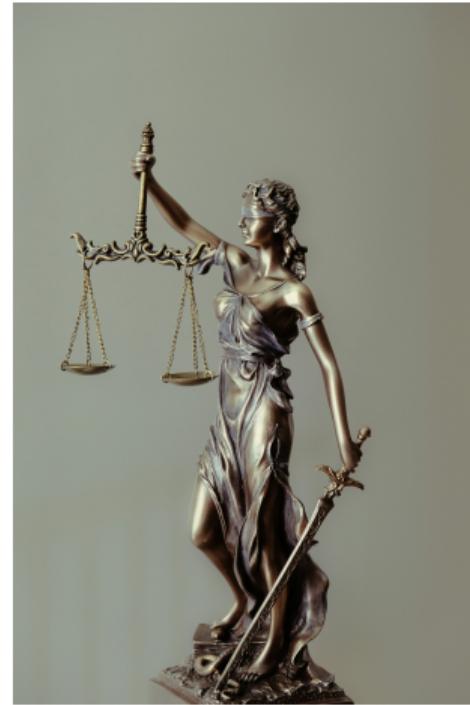


Fairness in Machine Learning

Juliet Fleischer

January 7, 2025



Ist die Entscheidung zur Festnahme fair?

- Trainingsdaten: SQF Daten von NYPD aus 2023 mit 16,924 Stopps
- Zielvariable: Festnahme (0 = keine Festnahme, 1 = Festnahme)
- **Protected Attribute:** Ethnie

race	prop
BLACK	58.61%
WHITE HISPANIC	20.32%
BLACK HISPANIC	10.13%
WHITE	5.48%
OTHER	2.67%
NA	2.79%



Agenda

1 Quellen von Bias

2 Gruppen Fairness

3 Individuelle Fairness

4 Fairness Methoden

5 Extra Folien

Woher kommt Bias?

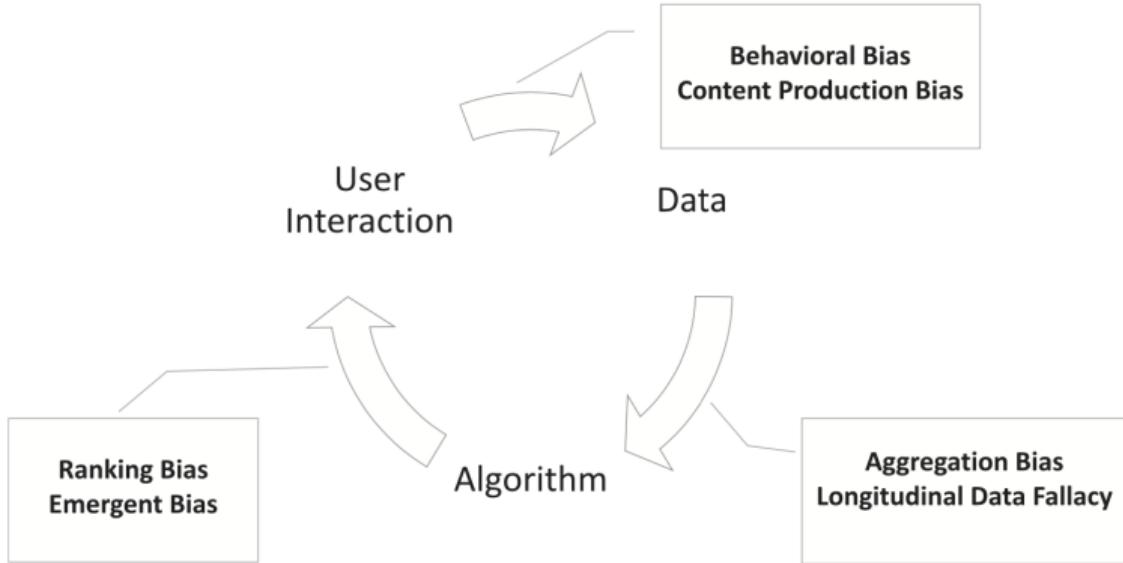


Figure: Quellen von Bias in Machine Learning Modellen [1]

Agenda

1 Quellen von Bias

2 Gruppen Fairness

3 Individuelle Fairness

4 Fairness Methoden

5 Extra Folien

Independence	Separation	Sufficiency
$\hat{Y} \perp A$	$\hat{Y} \perp A Y$	$Y \perp A \hat{Y}$

- Verständnis von Fairness: Personen erfahren aufgrund ihrer Gruppenzugehörigkeit Diskriminierung
- Vorhersageraten zwischen Gruppen sollen gleich sein

z.B. Demographic Parity [2]

$$P(\hat{Y} = 1|A = a) = P(\hat{Y} = 1|A = b)$$

Independence	Separation	Sufficiency
$\hat{Y} \perp A$	$\hat{Y} \perp A Y$	$Y \perp A \hat{Y}$

- Fokus auf gleichen Fehlerraten zwischen Gruppen

z.B. Predictive Equality

$$P(\hat{Y} = 1|A = a, Y = 0) = P(\hat{Y} = 1|A = b, Y = 0)$$

$$\Rightarrow \text{gleicheFPR} == \text{gleicheFNR}$$

	$Y = 0$	$Y = 1$
$\hat{Y} = 0$	TN	FN
$\hat{Y} = 1$	FP	TP

Independence	Separation	Sufficiency
$\hat{Y} \perp A$	$\hat{Y} \perp A Y$	$Y \perp A \hat{Y}$

- Positive/negative Vorhersage soll die selbe Bedeutung für alle Gruppen haben

z.B. Predictive Parity

$$P(Y = 1|A = a, \hat{Y} = 1) = P(Y = 1|A = b, \hat{Y} = 1)$$

	$Y = 0$	$Y = 1$
$\hat{Y} = 0$	TN	FN
$\hat{Y} = 1$	FP	TP

- Sufficiency nimmt Perspektive des Entscheidenden an
 - Separation gut, wenn Y objektive Wahrheit oder zuverlässig
 - Independence gut, wenn Form der Gleichheit erzwungen werden soll
- normalerweise nicht miteinander vereinbar

Agenda

- 1 Quellen von Bias
- 2 Gruppen Fairness
- 3 Individuelle Fairness
- 4 Fairness Methoden
- 5 Extra Folien

- Verständnis von Fairness: Fairness bedeutet, dass gleiche Personen gleich behandelt werden

① Fairness through Awareness (FTA)

- ▶ Lipschitz-Kriterium:

$$d(\hat{y}_i, \hat{y}_j) \leq \lambda d(x_i, x_j)$$

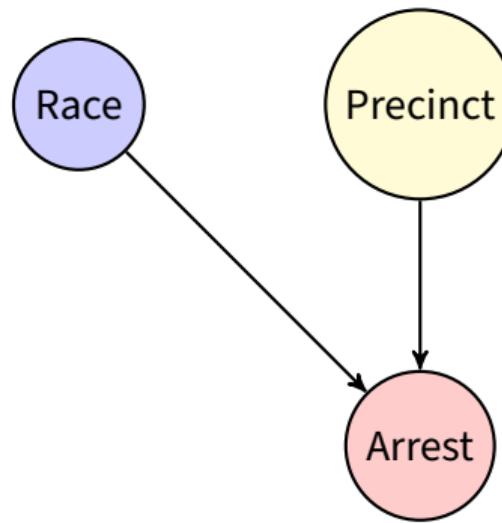
- ▶ Lässt sich in ein lineares Optimierungsproblem umformulieren
- ▶ Definition des Distanzmaßes d im Feature Space ist eine Herausforderung

② Fairness through Unawareness (FTU) = Blinding

- ▶ Vorgehensvorschrift: PA soll nicht im Entscheidungsprozess verwendet werden
- ▶ Keine eindeutige mathematische Definition, sondern verschiedene Ansätze zum Testen von FTU
- ▶ Problem der Proxis (Variablen, die mit PA hoch korreliert sind)

Ist die Gruppenzugehörigkeit die Grund für die Festnahme?

- kausale Definitionen



Race	Precinct	Prior Offenses	Arrest
Black	1	Yes	Yes
White	2	No	No
Hispanic	1	Yes	No
Asian	3	No	No

Agenda

1 Quellen von Bias

2 Gruppen Fairness

3 Individuelle Fairness

4 Fairness Methoden

5 Extra Folien

Wie sorgen wir für algorithmische Fairness?

- Preprocessing: Daten vor dem Training bearbeiten
z.B. (Re-)Sampling, Transformation
- Inprocessing: Trainingsprozess anpassen, Optimierungsproblem modifizieren
z.B. Regulaisierung
- Postprocessing: Vorhersagen nach dem Training bearbeiten
z.B. Thresholding
- Interpretable ML Methoden können hier auch sehr helfen!

Wie geht es weiter?

- binäre Klassifikation, ein PA ist simpelster Fall
- in Praxis eher mehrere PAs und vielfältige Aufgaben
→ regression, unsupervised learning, ...
- SQF Datensatz in Fairness Literatur in vielen weiteren Fragen untersucht

Fazit: Fairness ist ein komplexes Thema, das Zusammenarbeit vieler Disziplinen erfordert

-  **Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan.**
A Survey on Bias and Fairness in Machine Learning.
54(6):1–35.
-  **Sahil Verma and Julia Rubin.**
Fairness definitions explained.
In *Proceedings of the International Workshop on Software Fairness*, pages 1–7. ACM.

Agenda

1 Quellen von Bias

2 Gruppen Fairness

3 Individuelle Fairness

4 Fairness Methoden

5 Extra Folien

- Interaktives Tool:
<https://research.google.com/bigpicture/attacking-discrimination-in-ml/>
- Einführung in Fairness mit mlr3.fairness:
<https://journal.r-project.org/articles/RJ-2023-034/>
- Fairness and Machine Learning Buch: <https://fairmlbook.org/>

POC werden überproportional häufig gestoppt

Verteilung der Ethnie in NYC (2023)

<https://www.census.gov/quickfacts/newyorkcitynewyork>

Table: Verteilung der Ethnie in SQF Daten

race	prop
BLACK	58.61%
WHITE HISPANIC	20.32%
BLACK HISPANIC	10.13%
WHITE	5.48%
OTHER	2.67%
NA	2.79%