

# Fair Machine Learning

Juliet Fleischer

13. Januar 2025



# Ist die Vorhersage fair?

- Ziel: Kriminalitätsrate in NYC reduzieren

# Ist die Vorhersage fair?

- Ziel: Kriminalitätsrate in NYC reduzieren
- Zielvariable: Straftat begangen (0 = Nein, 1 = Ja)

# Ist die Vorhersage fair?

- Ziel: Kriminalitätsrate in NYC reduzieren
- Zielvariable: Straftat begangen (0 = Nein, 1 = Ja)
- Trainingsdaten: vergangene Polizeistops

# Ist die Vorhersage fair?

- Ziel: Kriminalitätsrate in NYC reduzieren
- Zielvariable: Straftat begangen (0 = Nein, 1 = Ja)
- Trainingsdaten: vergangene Polizeistops
- Problem: Daten könnten vergangene Diskriminierung widerspiegeln (racial profiling,...)  
⇒ Wie können wir sicherstellen, dass der Algorithmus gerecht entscheidet?

# Agenda

- 1 Gruppen Fairness
- 2 Individuelle Fairness
- 3 Fairness Methoden
- 4 Bias und die Feedback Loop
- 5 Extra Folien

# Agenda

- 1 Gruppen Fairness
- 2 Individuelle Fairness
- 3 Fairness Methoden
- 4 Bias und die Feedback Loop
- 5 Extra Folien

# Gleiche Vorhersageraten zwischen Gruppen sind fair

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

# Gleiche Vorhersageraten zwischen Gruppen sind fair

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

# Gleiche Vorhersageraten zwischen Gruppen sind fair

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
- Gruppenzugehörigkeit über **Protected Attribute (PA)** definiert

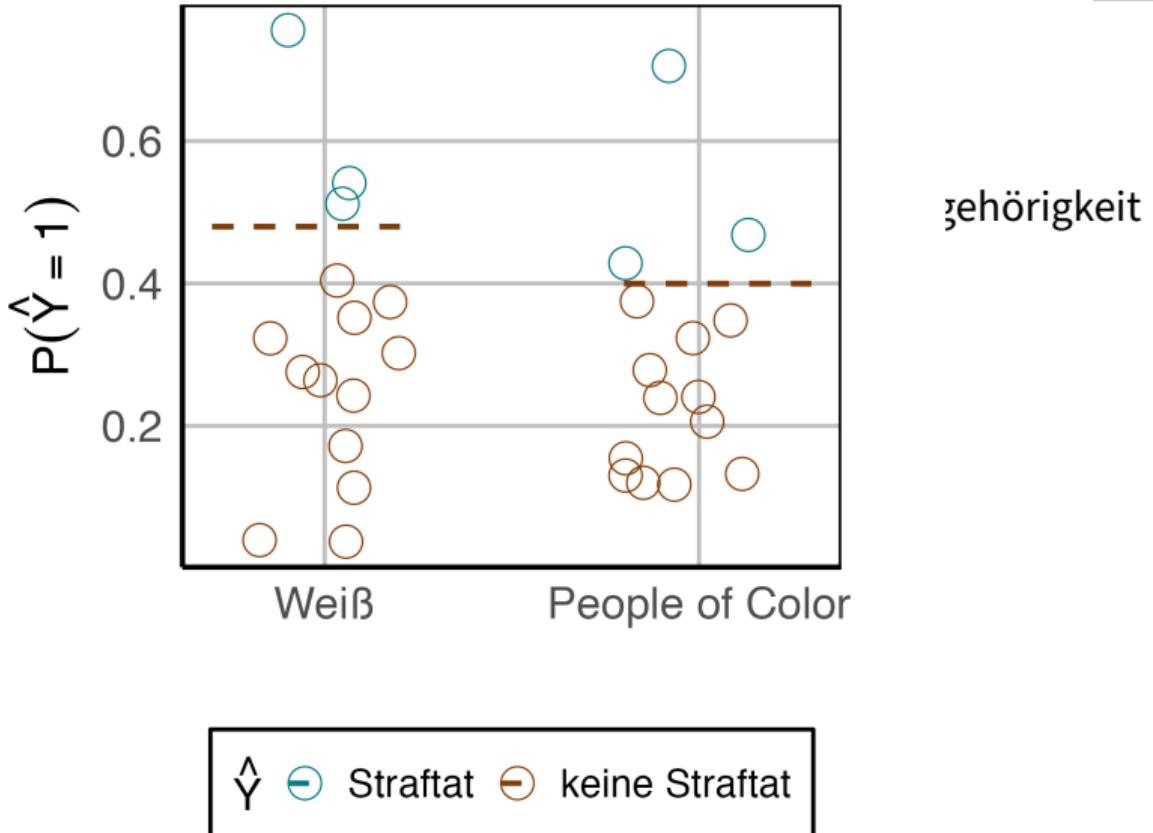
# Gleiche Vorhersageraten zwischen Gruppen sind fair

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
- Gruppenzugehörigkeit über **Protected Attribute (PA)** definiert
- Vorhersageraten zwischen Gruppen sollen gleich sein

# Gleiche Vorher...

- Verständnis von Diskriminierung
- Gruppenzugehörigkeit
- Vorhersagerat



# Gleiche Vorhersageraten zwischen Gruppen sind fair

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
- Gruppenzugehörigkeit über **Protected Attribute (PA)** definiert
- Vorhersageraten zwischen Gruppen sollen gleich sein
- z.B. Statistical Parity [7]

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

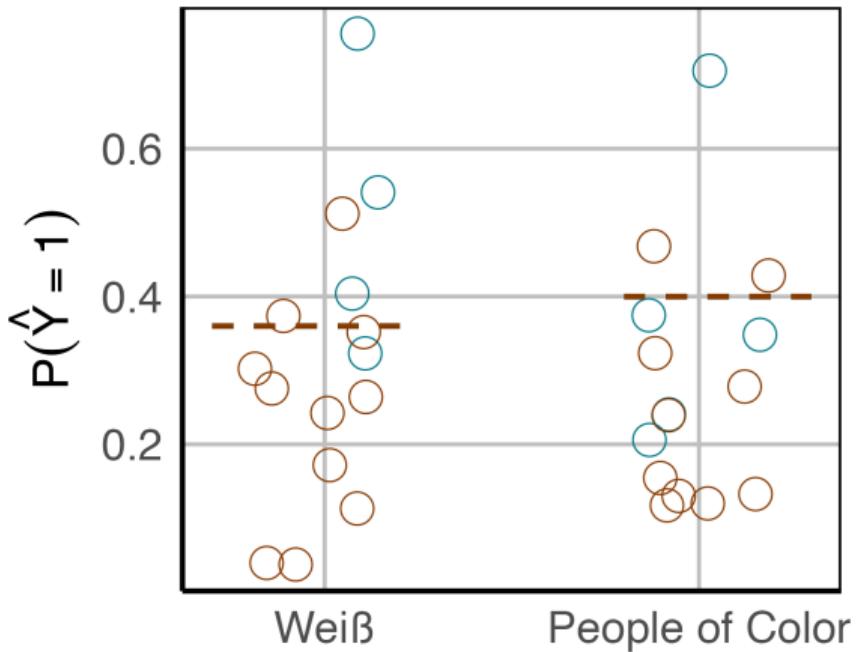
# Fairness anhand der Fehlermatrix konstruieren

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

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

- **Separation:** Fokus auf gleichen Fehlerraten zwischen Gruppen

- **Separation:** Fc  
Fehlerraten zw.



Y    Straftat    keine Straftat

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

- **Separation:** Fokus auf gleichen Fehlerraten zwischen Gruppen

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

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

- **Separation:** Fokus auf gleichen Fehlerraten zwischen Gruppen
- z.B. Predictive Equality [7]

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

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

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

- **Separation:** Fokus auf gleichen Fehlerraten zwischen Gruppen
- z.B. Predictive Equality [7]

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

- **Sufficiency:** Zuverlässigkeit der Vorhersage soll gleich sein

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

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

- **Separation:** Fokus auf gleichen Fehlerraten zwischen Gruppen
- z.B. Predictive Equality [7]

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

- **Sufficiency:** Zuverlässigkeit der Vorhersage soll gleich sein
- z.B. Predictive Parity [7]

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

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

# Gruppenmetriken im Überblick

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

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

- Zahlreiche Variationen von Gruppen-Metriken

z.B. Calibration  $P(Y = 1|A = a, S = s) = P(Y = 1|A = b, S = s)$

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

- Zahlreiche Variationen von Gruppen-Metriken  
z.B. Calibration  $P(Y = 1|A = a, S = s) = P(Y = 1|A = b, S = s)$
- Sufficiency nimmt Perspektive des Entscheidenden an [1]

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

- Zahlreiche Variationen von Gruppen-Metriken  
z.B. Calibration  $P(Y = 1|A = a, S = s) = P(Y = 1|A = b, S = s)$
- Sufficiency nimmt Perspektive des Entscheidenden an [1]
- Separation gut, wenn Y durch einen objektiv wahren Prozess entstanden ist [1]

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

- Zahlreiche Variationen von Gruppen-Metriken  
z.B. Calibration  $P(Y = 1|A = a, S = s) = P(Y = 1|A = b, S = s)$
- Sufficiency nimmt Perspektive des Entscheidenden an [1]
- Separation gut, wenn Y durch einen objektiv wahren Prozess entstanden ist [1]
- Independence gut, wenn Form der Gleichheit erzwungen werden soll [1]

⇒ normalerweise nicht miteinander vereinbar

# Agenda

- 1 Gruppen Fairness
- 2 Individuelle Fairness
- 3 Fairness Methoden
- 4 Bias und die Feedback Loop
- 5 Extra Folien

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

- Verständnis von Fairness: Fairness bedeutet, dass gleiche Personen gleich behandelt werden
- ① Fairness through Awareness [4] (FTA)

- Verständnis von Fairness: Fairness bedeutet, dass gleiche Personen gleich behandelt werden
- ① Fairness through Awareness [4] (FTA)
  - ▶ Lipschitz-Kriterium [1]:

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

- Verständnis von Fairness: Fairness bedeutet, dass gleiche Personen gleich behandelt werden
- ① Fairness through Awareness [4] (FTA)
  - ▶ Lipschitz-Kriterium [1]:
$$d_Y(\hat{y}_i, \hat{y}_j) \leq \lambda d_X(x_i, x_j)$$
  - ▶ Definition des Distanzmaßes  $d_X$  im Feature Space ist eine Herausforderung

- Verständnis von Fairness: Fairness bedeutet, dass gleiche Personen gleich behandelt werden
- ① Fairness through Awareness [4] (FTA)
  - ▶ Lipschitz-Kriterium [1]:
$$d_Y(\hat{y}_i, \hat{y}_j) \leq \lambda d_X(x_i, x_j)$$
  - ▶ Definition des Distanzmaßes  $d_X$  im Feature Space ist eine Herausforderung
- ② Fairness through Unawareness (FTU) = Blinding

- Verständnis von Fairness: Fairness bedeutet, dass gleiche Personen gleich behandelt werden
- ① Fairness through Awareness [4] (FTA)
  - ▶ Lipschitz-Kriterium [1]:
$$d_Y(\hat{y}_i, \hat{y}_j) \leq \lambda d_X(x_i, x_j)$$
  - ▶ Definition des Distanzmaßes  $d_X$  im Feature Space ist eine Herausforderung
- ② Fairness through Unawareness (FTU) = Blinding
  - ▶ Vorgehensvorschrift: PA soll nicht im Entscheidungsprozess verwendet werden

- Verständnis von Fairness: Fairness bedeutet, dass gleiche Personen gleich behandelt werden
- ① Fairness through Awareness [4] (FTA)
  - ▶ Lipschitz-Kriterium [1]:
$$d_Y(\hat{y}_i, \hat{y}_j) \leq \lambda d_X(x_i, x_j)$$
  - ▶ Definition des Distanzmaßes  $d_X$  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

- Verständnis von Fairness: Fairness bedeutet, dass gleiche Personen gleich behandelt werden
- ① Fairness through Awareness [4] (FTA)
  - ▶ Lipschitz-Kriterium [1]:
$$d_Y(\hat{y}_i, \hat{y}_j) \leq \lambda d_X(x_i, x_j)$$
  - ▶ Definition des Distanzmaßes  $d_X$  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)

# Agenda

- 1 Gruppen Fairness
- 2 Individuelle Fairness
- 3 Fairness Methoden
- 4 Bias und die Feedback Loop
- 5 Extra Folien

# Wie sorgen wir für algorithmische Fairness?

- Preprocessing [2]: Daten vor dem Training bearbeiten z.B. (Re-)Sampling, Transformation

# Wie sorgen wir für algorithmische Fairness?

- Preprocessing [2]: Daten vor dem Training bearbeiten  
z.B. (Re-)Sampling, Transformation
- Inprocessing [2]: Trainingsprozess anpassen, Optimierungsproblem modifizieren  
z.B. Regulaisierung

# Wie sorgen wir für algorithmische Fairness?

- Preprocessing [2]: Daten vor dem Training bearbeiten  
z.B. (Re-)Sampling, Transformation
- Inprocessing [2]: Trainingsprozess anpassen, Optimierungsproblem modifizieren  
z.B. Regulaisierung
- Postprocessing [2]:Vorhersagen nach dem Training bearbeiten  
z.B. Thresholding

- Preprocessing [2]: Daten vor dem Training bearbeiten  
z.B. (Re-)Sampling, Transformation
- Inprocessing [2]: Trainingsprozess anpassen, Optimierungsproblem modifizieren  
z.B. Regulaisierung
- Postprocessing [2]:Vorhersagen nach dem Training bearbeiten  
z.B. Thresholding
- Interpretable ML Methoden können hier auch sehr helfen!

# Agenda

- 1 Gruppen Fairness
- 2 Individuelle Fairness
- 3 Fairness Methoden
- 4 Bias und die Feedback Loop
- 5 Extra Folien

# Woher kommt Bias?

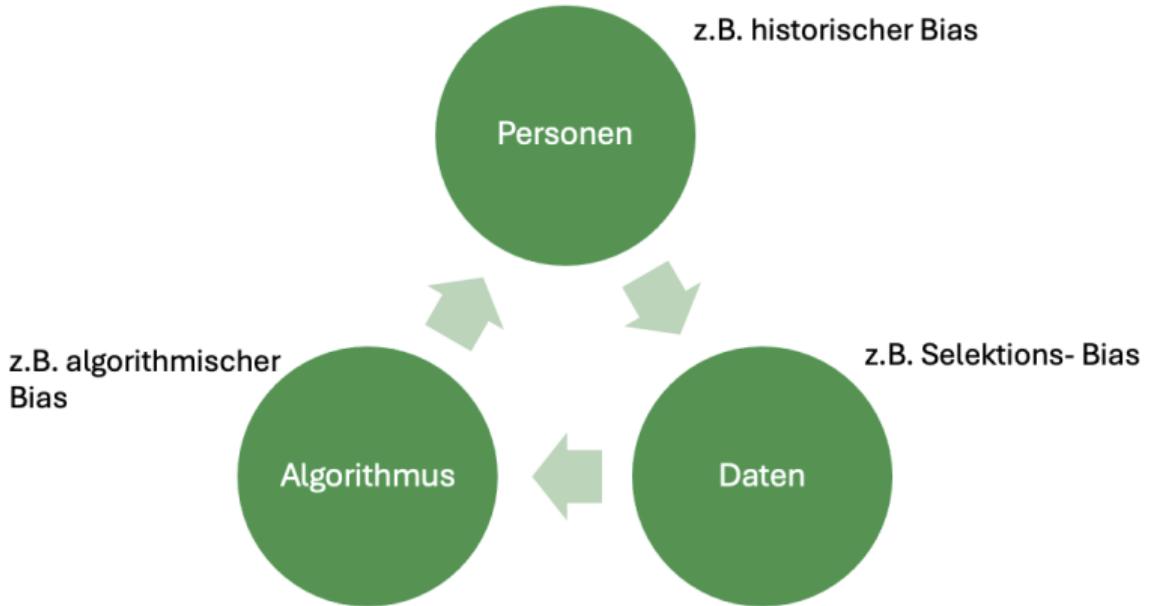


Abbildung: Quellen von Bias in der Daten, Nutzer, Algorithmus Feedback Loop [6]

# Wie geht es weiter?

- Binäre Klassifikation, ein PA ist simpelster Fall

- Binäre Klassifikation, ein PA ist simpelster Fall
- In Praxis eher mehrere PAs und vielfältige Aufgaben  
→ regression, unsupervised learning, ...

# Wie geht es weiter?

- Binäre Klassifikation, ein PA ist simpelster Fall
- In Praxis eher mehrere PAs und vielfältige Aufgaben  
→ regression, unsupervised learning, ...
- Es gibt (noch) nicht, die **eine** Definition von Fairness

- [1] Alessandro Castelnovo u. a. “A Clarification of the Nuances in the Fairness Metrics Landscape”. In: *Scientific Reports* 12.1 (März 2022), S. 4209. ISSN: 2045-2322. DOI: 10.1038/s41598-022-07939-1. (Besucht am 23. 12. 2024).
- [2] Simon Caton und Christian Haas. “Fairness in Machine Learning: A Survey”. In: *ACM Computing Surveys* 56.7 (Juli 2024), S. 1–38. ISSN: 0360-0300, 1557-7341. DOI: 10.1145/3616865. (Besucht am 23. 12. 2024).
- [3] Sam Corbett-Davies u. a. “The Measure and Mismeasure of Fairness”. In: () .
- [4] Cynthia Dwork u. a. “Fairness through Awareness”. In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. Cambridge Massachusetts: ACM, Jan. 2012, S. 214–226. ISBN: 978-1-4503-1115-1. DOI: 10.1145/2090236.2090255. (Besucht am 29. 12. 2024).
- [5] Matt J Kusner u. a. “Counterfactual Fairness”. In: () .

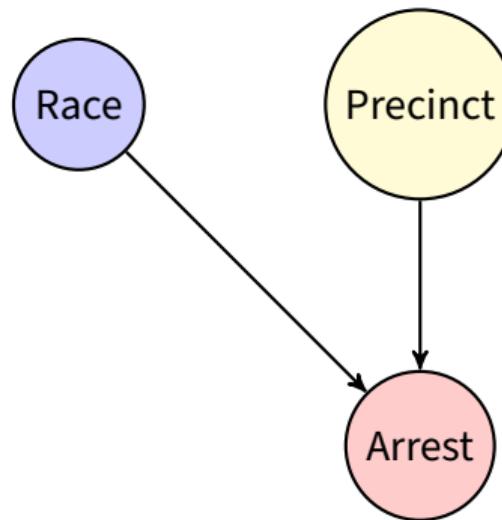
- [6] Ninareh Mehrabi u. a. “A Survey on Bias and Fairness in Machine Learning”. In: *ACM Computing Surveys* 54.6 (Juli 2022), S. 1–35. ISSN: 0360-0300, 1557-7341. DOI: 10.1145/3457607. (Besucht am 07. 01. 2025).
- [7] Sahil Verma und Julia Rubin. “Fairness Definitions Explained”. In: *Proceedings of the International Workshop on Software Fairness*. Gothenburg Sweden: ACM, Mai 2018, S. 1–7. ISBN: 978-1-4503-5746-3. DOI: 10.1145/3194770.3194776. (Besucht am 16. 11. 2024).
- [8] Muhammad Bilal Zafar u. a. “From Parity to Preference-based Notions of Fairness in Classification”. In: *Advances in Neural Information Processing Systems*. Bd. 30. Curran Associates, Inc., 2017. (Besucht am 29. 12. 2024).

# Agenda

- 1 Gruppen Fairness
- 2 Individuelle Fairness
- 3 Fairness Methoden
- 4 Bias und die Feedback Loop
- 5 Extra Folien

- Gruppen Metriken spiegeln einfach Verteilung von  $Y, A, \hat{Y}, X$  in den Daten wider [3]
- Accuracy und Fairness Trade-Off
- Folgen von Fairness Interventionen nicht sicher - profitiert die geschützte Gruppe?

# Ist die Gruppenzugehörigkeit der Grund für die Festnahme? - kausale Definitionen



- Gruppen Fairness: FACE, FACT (on average or on conditional average level) [8]
- Individuelle Fairness: counterfactual fairness, path-based fairness [5]

- 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>

Tabelle: 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%