Karlsruher Institut für Technologie

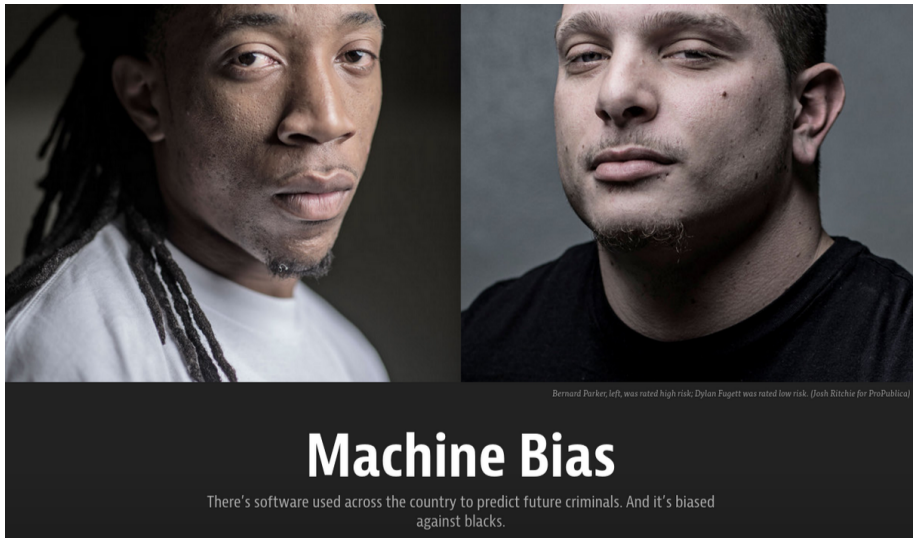
Diskriminierung aus Sicht v. Philosophie & Informatik

Seminar im Sommersemester 2021

Auftaktveranstaltung | 05. Mai 2021

INSTITUT FÜR INFORMATIONSSICHERHEIT UND VERLÄSSLICHKEIT & INSTITUT FÜR TECHNIKZUKÜNFTE

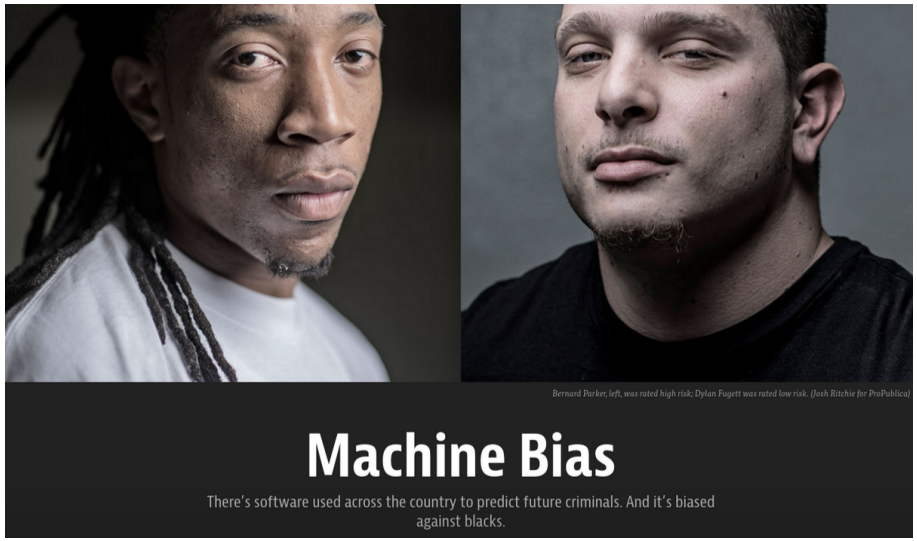
Motivation: Das COMPAS-Beispiel[1]



- Prognosesoftware zur Rückfallwahrscheinlichkeit von Straftäter:innen
- Basiert auf Antworten zu 137 Fragen durch Angeklagte und Strafregister
- Fand in vielen Gerichtsurteilen der USA für mehrere Jahre Verwendung

[1] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

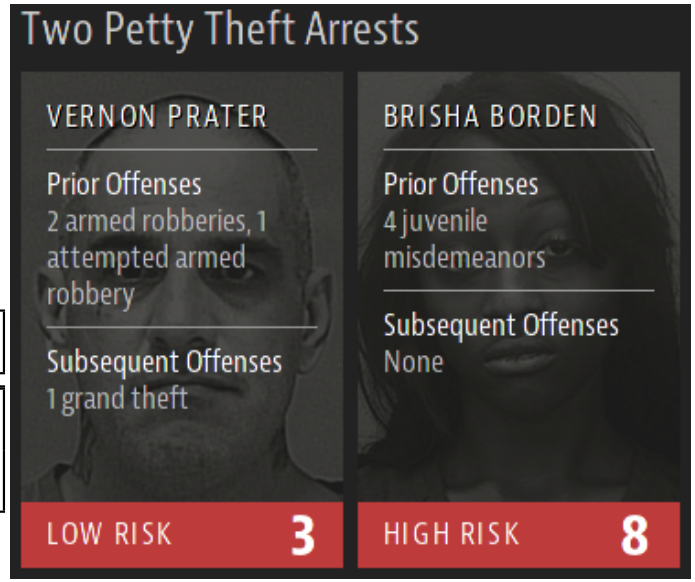
Motivation: Das COMPAS-Beispiel[1]



- Prognosesoftware zur Rückfallwahrscheinlichkeit von Straftäter:innen
- Basiert auf Antworten zu 137 Fragen durch Angeklagte und Strafregister
- Fand in vielen Gerichtsurteilen der USA für mehrere Jahre Verwendung

- Attestierte Präzision von 68 Prozent (69% für Weiße, 67% für Afro-Amerik.)
- Jedoch . . .

	Weiß	Afro-Am.
„Hohes Risiko“, kein Rückfall	23,5 %	44,9 %
„Geringes Risiko“, Rückfall	47,7 %	28,0 %



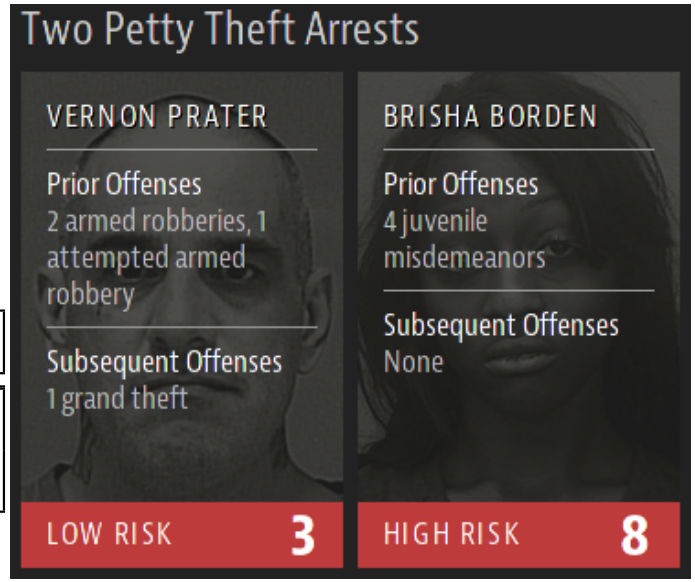
[1] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Motivation: Das COMPAS-Beispiel[1]

- Prognosesoftware zur Rückfallwahrscheinlichkeit von Straftäter:innen
- Basiert auf Antworten zu 137 Fragen durch Angeklagte und Strafregister
- Fand in vielen Gerichtsurteilen der USA für mehrere Jahre Verwendung

- Atte n 68 Prozent (69% ir Afro-Amerik.)
- Jedc

	Weiß	Afro-Am.
„Hohes Risiko“, kein Rückfall	23,5 %	44,9 %
„Geringes Risiko“, Rückfall	47,7 %	28,0 %



[1] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Motivation: Themenbereich

Diskriminierung durch (maschinell-gelernte) Entscheidungsverfahren

Motivation: Themenbereich

Diskriminierung durch (maschinell-gelernte) Entscheidungsverfahren

Hierzu Fragestellungen an der Verbindung zwischen . . .

- **Praktischer Philosophie**: Ethik
- und **Theoretischer Informatik**: Formale Logik

Motivation: Themenbereich

Diskriminierung durch (maschinell-gelernte) Entscheidungsverfahren

Hierzu Fragestellungen an der Verbindung zwischen . . .

- **Praktischer Philosophie**: Ethik
- und **Theoretischer Informatik**: Formale Logik

Beispielhafte Fragestellungen:

- Unter welchen Bedingungen spricht man von Diskriminierung?
- Sind faire Entscheidungen überhaupt möglich bzw. nützlich?
- Wie kann man sich sicher sein, dass ein Algorithmus fair handelt?
- Was bedeutet Fairness genau? Ist das eindeutig?
- Wie kann man Unfairness abmildern?

Ablauf und Organisatorisches

1. Recherche auf Basis von Einstiegspapieren und regelm. Treffen
   $\Rightarrow$ Verstehen und Eingrenzen des Themas
2. Kurze Vorstellung der Gliederung im Plenum
3. Vortragsplanung und Erstellen des Hauptvortrags
4. Ab 10. Juni 4-7 wöchentl. Termine zu ausgewählten Papieren
5. Am 12./15. Juli Block mit Gastreferenten, Vorträgen u. Diskussionen
6. Schriftliche Ausarbeitung (ca. 10 – 15 Seiten LNCS)

Ablauf und Organisatorisches

1. Recherche auf Basis von Einstiegspapieren und regelm. Treffen
   $\Rightarrow$ Verstehen und Eingrenzen des Themas
2. Kurze Vorstellung der Gliederung im Plenum
3. Vortragsplanung und Erstellen des Hauptvortrags
4. Ab 10. Juni 4-7 wöchentl. Termine zu ausgewählten Papieren
5. Am 12./15. Juli Block mit Gastreferenten, Vorträgen u. Diskussionen
6. Schriftliche Ausarbeitung (ca. 10 – 15 Seiten LNCS)

Kurze Kennenlernrunde

- Wer sind Sie & was studieren bzw. in welchem Gebiet arbeiten Sie?
- Was ist Ihre Motivation für das Seminar?
- Ggfs.: Wie möchten Sie das Seminar anrechnen? (Inf./Phil./SQ)

Angebotene Themen

#	Thema
1.	Limitations of common fairness notions
2.	Practical trade-offs of fairness notions
3.	Unveiling unfairness from datasets
4.	Can we make algorithms neutral?
5.	The relevance of false positive/negative rates
6.	The proxy problem and intransparent discrimination

Angebotene Themen

#	Thema
1.	Limitations of common fairness notions
2.	Practical trade-offs of fairness notions
3.	Unveiling unfairness from datasets
4.	Can we make algorithms neutral?
5.	The relevance of false positive/negative rates
6.	The proxy problem and intransparent discrimination

Weiterer Ablauf

- Themenwahl bis Freitag 11:00 Uhr an `kirsten@kit.edu`
- Zuteilung bis Montag, dann selbstst. zeitnah Betreuer kontaktieren