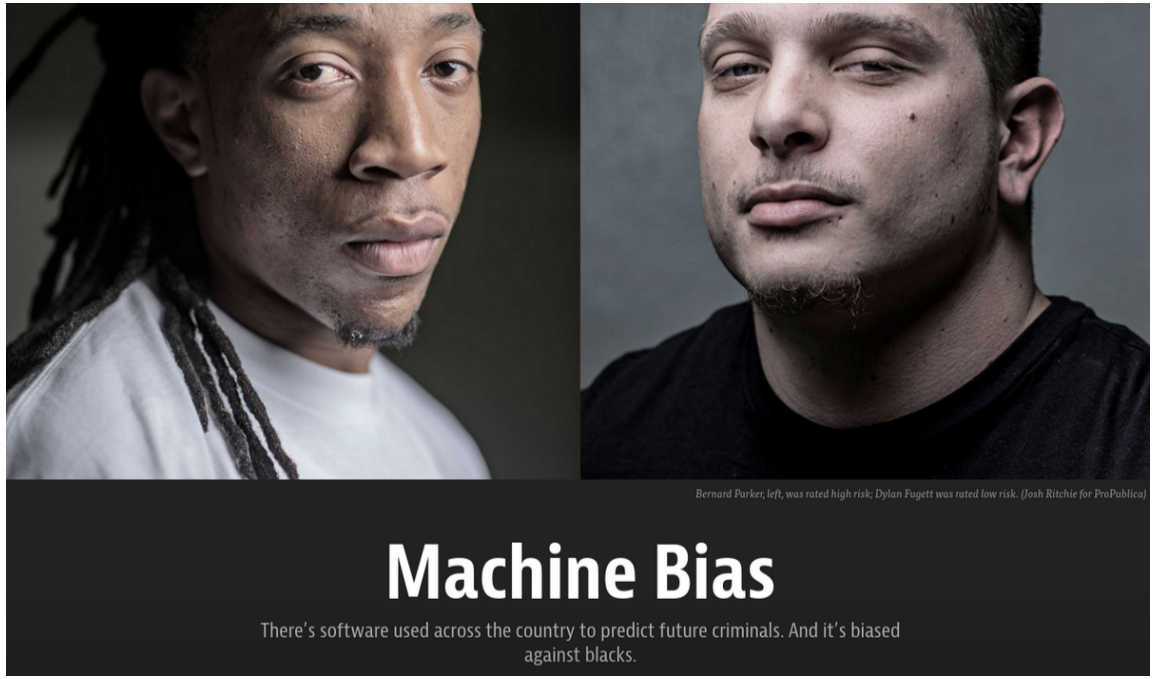
Fairness und Diskriminierungsfreiheit aus Sicht von Philosophie und Informatik

Seminar

Prof. Bernhard Beckert und Prof. Michael Schefczyk | 25. April 2025

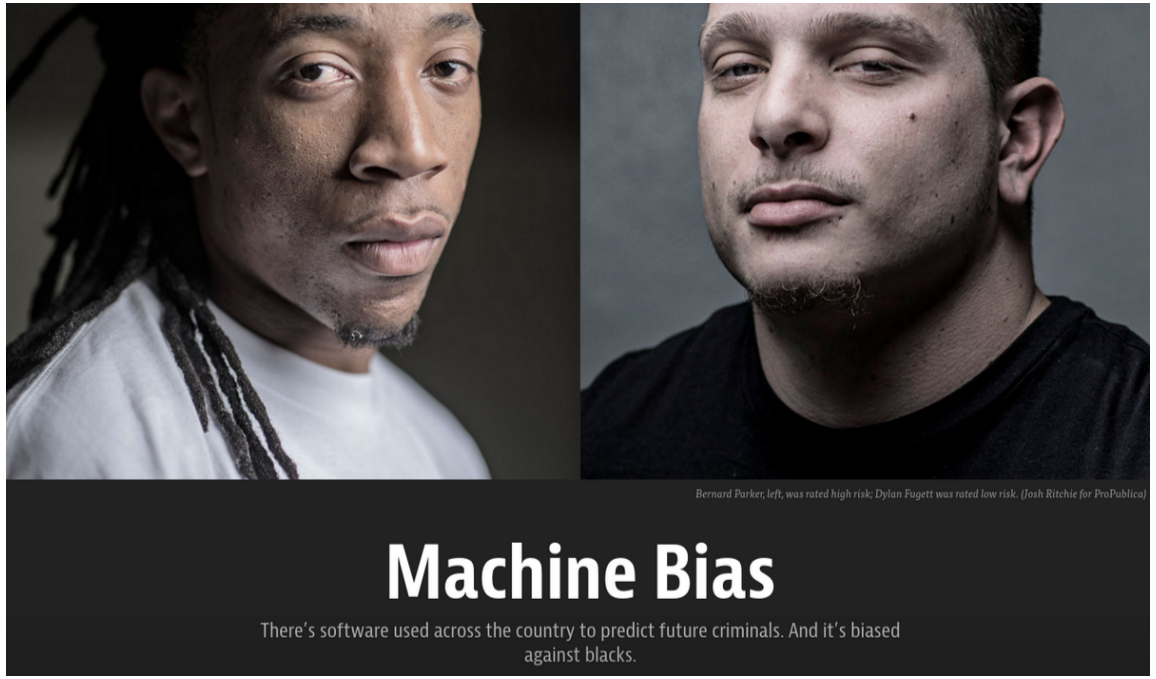
Motivation: Das COMPAS-Beispiel[1]



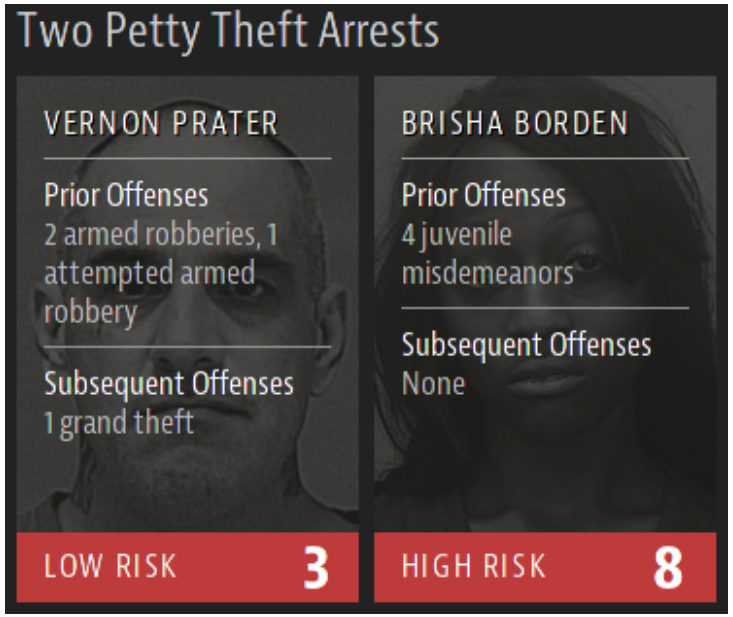
- Prognosesoftware zur Rückfallwahrscheinlichkeit von Straftäter:innen
- Basiert auf Antworten zu 137 Fragen durch Angeklagte und Strafregister
- Fand in vielen Gerichtsurteilen der USA für mehrere Jahre Verwendung

[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

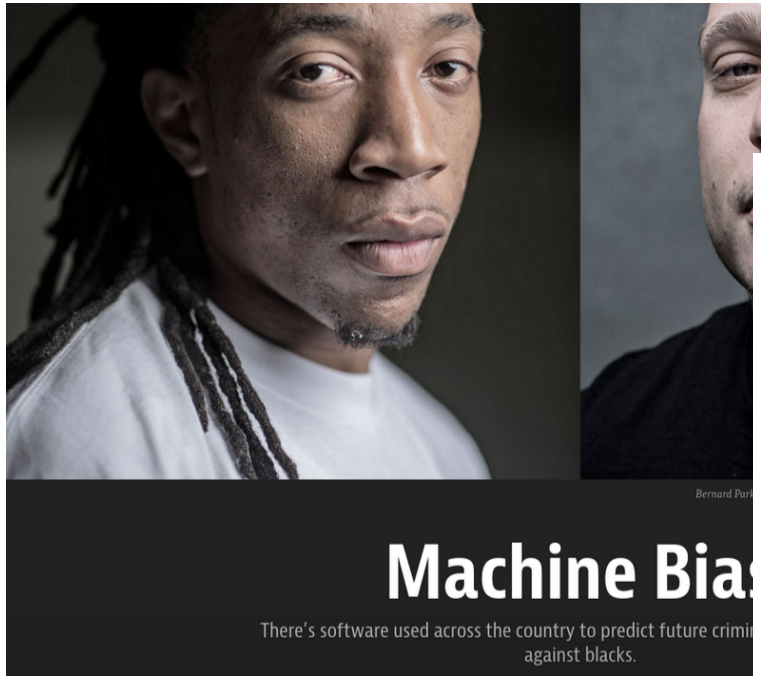


- Attestierte Präzision von 68 Prozent (69% für Weiße, 67% für Schwarze)
- Jedoch …

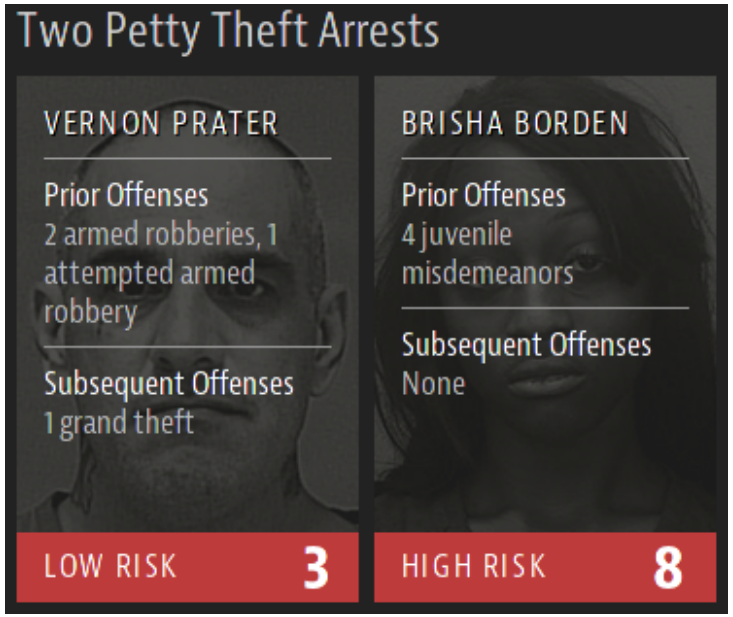
	Weiß	Schwarz
„Hohes Risiko“, kein Rückfall	23,5 %	44,9 %
„Geringes Risiko“, Rückfall	47,7 %	28,0 %

[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



- Attestierte Präzision von 68 Pr
  67% für Schwarze)
- Jedoch …

	Weiß	Schwarz
„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 sprechen wir von Diskriminierung?
- Sind faire Entscheidungen überhaupt möglich oder nützlich?
- Wie können wir uns sicher sein, dass ein Algorithmus fair handelt?
- Was bedeutet Fairness genau? Ist das eindeutig?
- Wie können wir Unfairness abmildern?

Zeitplan

25.04.2025	**Kickoff**
04.06.2025 **09.45 – 11.15 Uhr**	**Diskussion zu Kapitel 5 des Fair ML Book** `https://fairmlbook.org/causal.html`
16.07.2025 **09.45 – 13.00 Uhr**	**Impossibility (Themen 1–3)** **Equalized Odds and Pred. Parity | Yet Another Impossibility | Causal Hierarchy**
30.07.2025 **09.45 – 13.00 Uhr**	**Decomposing & Learning Causal Effects (Themen 4–6)** **Causal Explanation Formula | Path-Specific CF | Neural-Causal Connection**

Zeitplan

25.04.2025	**Kickoff**
04.06.2025 **09.45 – 11.15 Uhr**	**Diskussion zu Kapitel 5 des Fair ML Book** `https://fairmlbook.org/causal.html`
16.07.2025 **09.45 – 13.00 Uhr**	**Impossibility (Themen 1–3)** **Equalized Odds and Pred. Parity | Yet Another Impossibility | Causal Hierarchy**
30.07.2025 **09.45 – 13.00 Uhr**	**Decomposing & Learning Causal Effects (Themen 4–6)** **Causal Explanation Formula | Path-Specific CF | Neural-Causal Connection**

Termine für Blocksitzungen

1. 90 Minuten am 02. Juni nachmittags oder 04. Juni vormittags?
2. 180 Minuten am 07., 09., 14. oder vormittags am 16. Juli?
3. 180 Minuten am 28. Juli vormittags oder 30. Juli vormittags?

Vorgehen

1. Lesen: Kapitel 5 Fair ML Buch
2. Gemeinsame Diskussion beim ersten Termin
3. Einarbeitung in eigenes Thema (ggf. weitere Literaturrecherche)
4. Regelmäßige Besprechung mit Betreuern
5. Bis zum Blocktermin:
   - Vorbereitung Vortrag (30 Minuten)
   - Vorbereitung Diskussion (30 Minuten)
6. Im Anschluss: Schriftliche Ausarbeitung

Vorgehen

1. Lesen: Kapitel 5 Fair ML Buch
2. Gemeinsame Diskussion beim ersten Termin
3. Einarbeitung in eigenes Thema (ggf. weitere Literaturrecherche)
4. Regelmäßige Besprechung mit Betreuern
5. Bis zum Blocktermin:
   - Vorbereitung Vortrag (30 Minuten)
   - Vorbereitung Diskussion (30 Minuten)
6. Im Anschluss: Schriftliche Ausarbeitung

Erwartungen

- Sorgfältige, eigenständige Einarbeitung ins Thema
- Aufbereitung des Themas für Personen die Papiere nicht gelesen haben
- Aktive Teilnahme an Diskussionen

Themen

Betreuer	Studierende(r)	Thema
Block 1: (Im)possibility		
Michael Kirsten und Michael Schefczyk		Compatibility of Equalized Odds and Predictive Parity
Samuel Teuber		Yet Another Impossibility Theorem in Algorithmic Fairness
Samuel Teuber		On Pearl's Hierarchy and the Foundations of Causal Inference
Block 2: Decomposing and Learning Causal Effects		
Jonas Schiffl		Fairness in Decision-Making — The Causal Explanation Formula
Philipp Kern		Path-Specific Counterfactual Fairness
Jonas Klamroth*		The Causal-Neural Connection: Expressiveness, Learnability, and Inference

Bitte kontaktieren Sie Ihren jeweiligen Betreuer für eine erste Besprechung.

*Backup: Jonas Schiffl

Nutzung von Generativer KI

“In jedem Fall bleiben die Studierenden für ihre Arbeit verantwortlich. Dies gilt auch für die Teile ihrer Arbeit, die mit Hilfe von KI erstellt oder von ihr beeinflusst wurden.”

`https://www.informatik.kit.edu/downloads/studium/Leitfaden_Generative_KI_Informatik.pdf`

- Als Hilfswerkzeug in Ordnung (bspw. Grammatik, Übersetzung etc.)
- Sie sind für die Resultate verantwortlich
- Im Zweifelsfall mit Betreuer abklären

Bewertung

- Präsentation: 60%
- Seminararbeit: 30%
- Aktive Teilnahme an Diskussionen: 10%

Die Präsentation und Diskussionsleitung, die Abgabe der Seminararbeit, sowie die Diskussionsteilnahme sind alle zwingend erforderlich für das Bestehen des Kurses.