Karlsruher Institut für Technologie

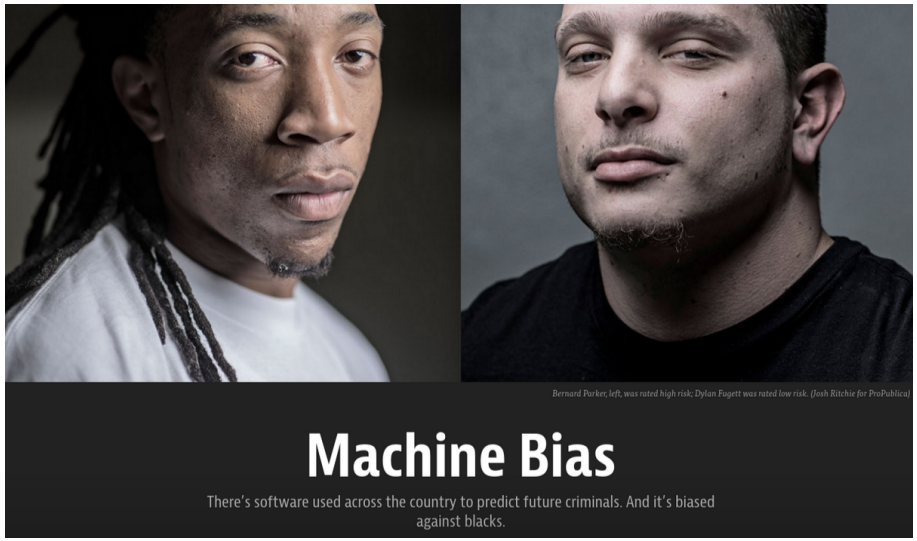
Fairness und Diskriminierungsfreiheit aus Sicht von Ethik und Informatik

Seminar im Sommersemester 2019

Kick-Off-Veranstaltung | 24. April 2019

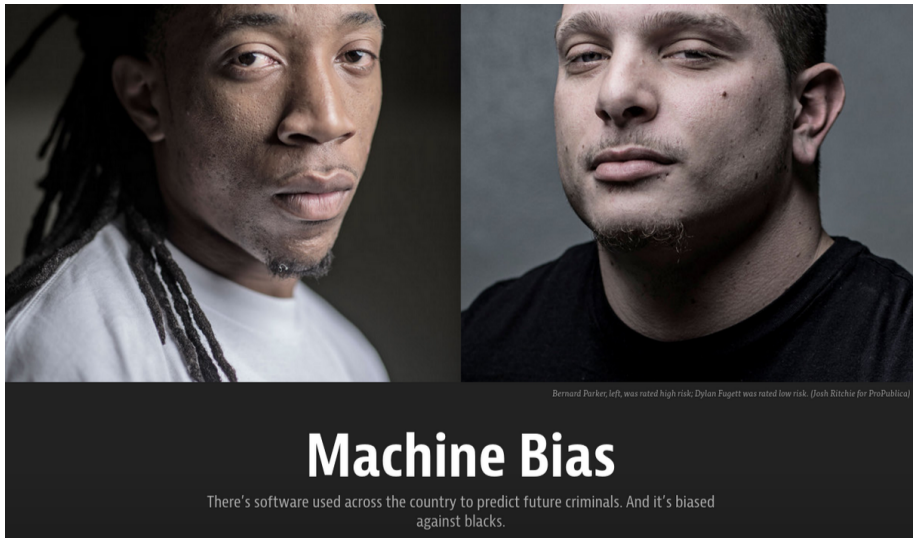
INSTITUT FÜR THEORETISCHE INFORMATIK & INSTITUT FÜR PHILOSOPHIE

Motivation: Ein Beispiel



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

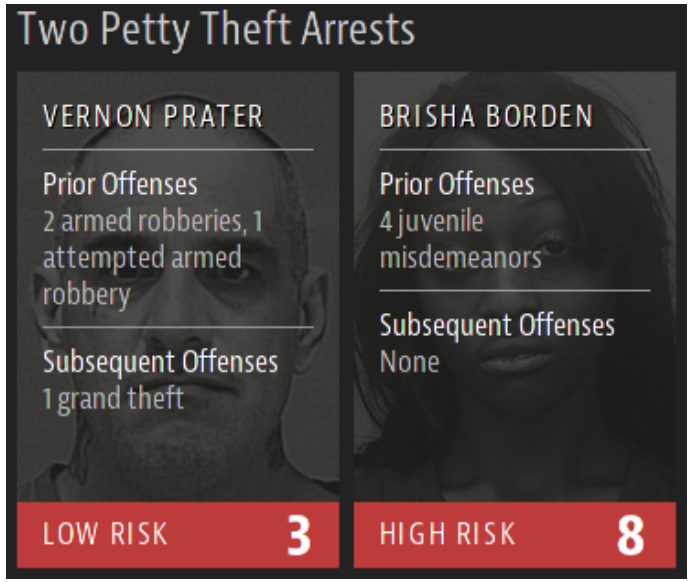
Motivation: Ein Beispiel



- Prognosesoftware zur Rückfallwahrscheinlichkeit von Straftätern
- 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 %



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 & Kryptographie

Motivation: Themenbereich

Diskriminierung durch (maschinell-gelernte) Entscheidungsverfahren

Hierzu Fragestellungen an der Verbindung zwischen . . .

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

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?

Zeitplan

1. Selbstständige Literaturrecherche auf Basis von Einstiegspapieren unter regelmäßiger Besprechung mit Betreuer
2. Verstehen und Eingrenzung des Themas für Präsentation
3. Kurze Vorstellung der Gliederung im Plenum
4. Planung des Seminarvortrags & evtl. vertiefte Recherche
5. Erstellen der Folien
6. Vortrag (30 Min. Vortrag + 15 Min. Diskussion)
7. Schriftliche Ausarbeitung (ca. 10 – 15 Seiten LNCS)

Zeitplan

1. Selbstständige Literaturrecherche auf Basis von Einstiegspapieren unter regelmäßiger Besprechung mit Betreuer
2. Verstehen und Eingrenzung des Themas für Präsentation
3. Kurze Vorstellung der Gliederung im Plenum $\Rightarrow$ **17.06.2019**
4. Planung des Seminarvortrags & evtl. vertiefte Recherche
5. Erstellen der Folien
6. Vortrag (30 Min. Vortrag + 15 Min. Diskussion) $\Rightarrow$ **26. & 29.07.2019**
7. Schriftliche Ausarbeitung (ca. 10 – 15 Seiten LNCS) $\Rightarrow$ **30.09.2019**

Angebotene Themen

#	Thema	Betreuer	Studieng.
1.	Discrimination in the Philosophical Debate	Prof. Schefczyk	EUKLID/SQ
2.	Discrimination as a Legal Term	Prof. Schefczyk	EUKLID/SQ
3.	Racial Profiling (4 topics)	Prof. Schefczyk	EUKLID/SQ
4.	Examples and Reasons for Unwanted Discrimination	Prof. Beckert	INF
5.	Techniques to Discover and Evaluate Unfairness	Michael Kirsten	INF
6.	Formalizing Fairness and Non-Discrimination	Prof. Beckert	INF
7.	Process, Outcome and Counterfactual Fairness	Jonas Schiffl	INF
8.	Trade-Offs and the Cost of Fairness	Jonas Schiffl	INF
9.	Fairness and Social Equality (Economic Models)	Michael Kirsten	INF
10.	Causal Reasoning for Fairness	Michael Kirsten	INF
11.	Approximate Fair Treatment and Sanitizing Classifiers	Alexander Koch	INF
12.	Composition of Approximate Fair Treatment	Alexander Koch	INF

Angebotene Themen

#	Thema	Betreuer	Studieng.
1.	Discrimination in the Philosophical Debate	Prof. Schefczyk	EUKLID/SQ
2.	Discrimination as a Legal Term	Prof. Schefczyk	EUKLID/SQ
3.	Racial Profiling (4 topics)	Prof. Schefczyk	EUKLID/SQ
4.	Examples and Reasons for Unwanted Discrimination	Prof. Beckert	INF
5.	Techniques to Discover and Evaluate Unfairness	Michael Kirsten	INF
6.	Formalizing Fairness and Non-Discrimination	Prof. Beckert	INF
7.	Process, Outcome and Counterfactual Fairness	Jonas Schiffl	INF
8.	Trade-Offs and the Cost of Fairness	Jonas Schiffl	INF
9.	Fairness and Social Equality (Economic Models)	Michael Kirsten	INF
10.	Causal Reasoning for Fairness	Michael Kirsten	INF
11.	Approximate Fair Treatment and Sanitizing Classifiers	Alexander Koch	INF
12.	Composition of Approximate Fair Treatment	Alexander Koch	INF

Weiterer Ablauf:

- Themenwahl bis morgen 18:30 Uhr an `kirsten@kit.edu`
- Zuteilung bis Freitag, dann zeitnah Betreuer kontaktieren