Fairness in KI-Systemen zur Beseitigung von Voreingenommenheit

INSTITUT POSITIVE COMPUTING

HRW

Warum Fairness?

Heutzutage werden viele unterschiedliche Bereiche durch KI-Systeme unterstützt oder gar automatisiert. KI-Systeme sind jedoch nur dann eine wertvolle Unterstützung, wenn sie auch korrekt verwendet werden. Das Problem hierbei besteht darin, dass der Mensch die KI-Systeme mit immensen Daten trainiert und das System die Objektivität innerhalb der Daten erlernt. Dies führt in den meisten Fällen zu einem ungewollten Bias und Fehleinschätzungen, sodass häufig der Gedanke der Fairness verletzt wird. Folgende unfaire KI-Systeme lösten prekäre Debatten aus:

- Google Translator (Abb. 1)
- Google Photos (Abb. 2)
- Amazons Recruiting-Engine
- COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) (Abb. 3)

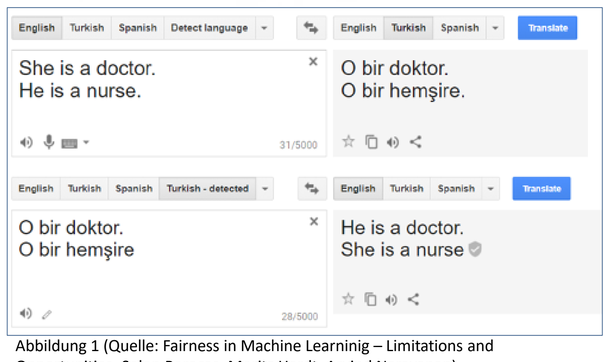


Abbildung 1 (Quelle: Fairness in Machine Learninig – Limitations and Opportunities, Solon Barocas, Moritz Hardt, Arvind Narayanan)

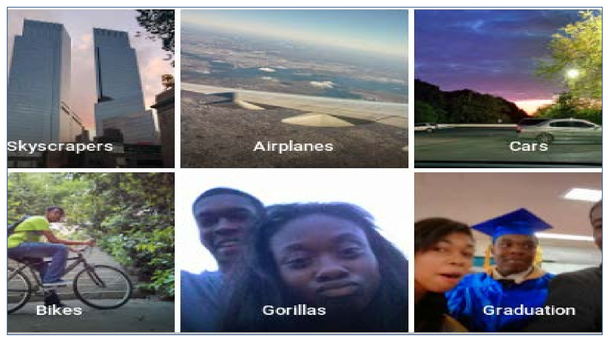


Abbildung 2 (Quelle: https://twitter.com/jackyalcine/status/615329515909156865)

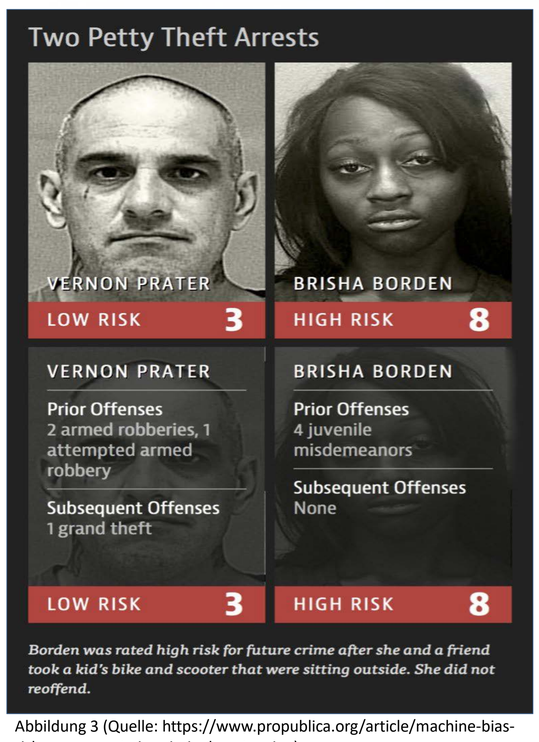


Abbildung 3 (Quelle: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing)

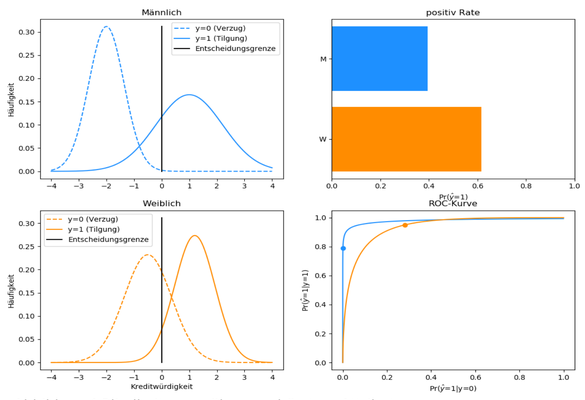


Abbildung 4 Blindheit gegenüber geschütztem Attribut

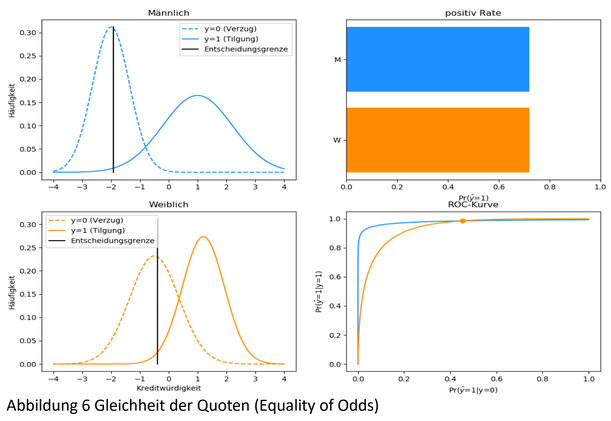


Abbildung 6 Gleichheit der Quoten (Equality of Odds)

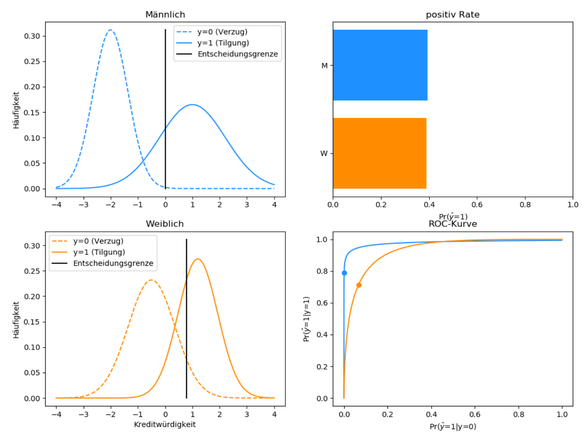


Abbildung 5 Demographische Parität

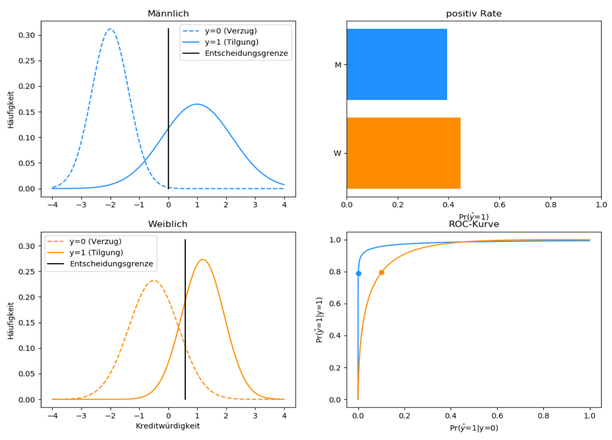


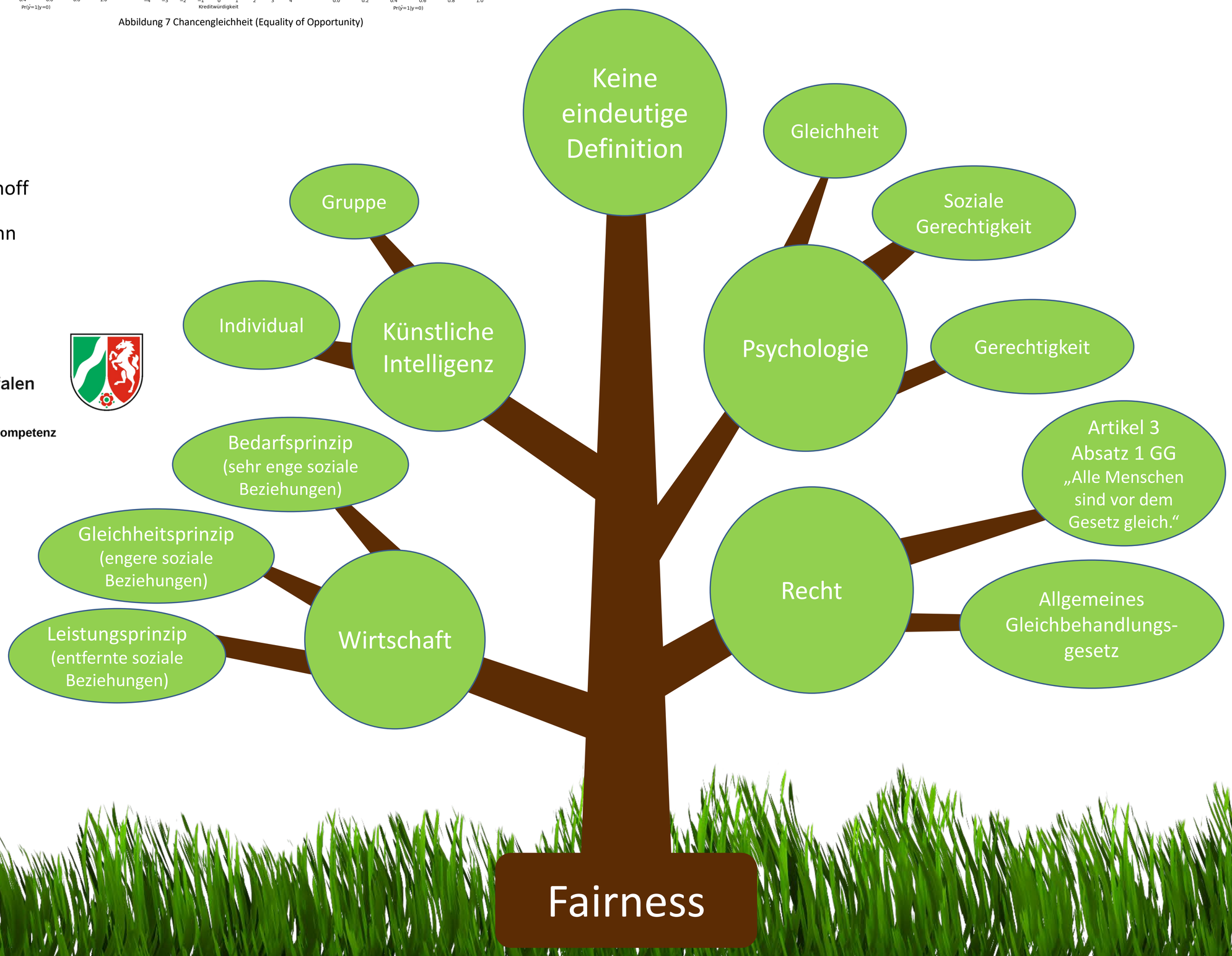
Abbildung 7 Chancengleichheit (Equality of Opportunity)

Projektidee

Bevor KI-Systeme fairer gestaltet werden können und somit die Beseitigung von Voreingenommenheit erfolgt, müssen unter anderem grundlegende Fragen wie *„Was ist Fairness?“*, *„Wer definiert Fairness?“*, und *„Wie kann Fairness mathematisch definiert werden?“ (Abb. 4,5,6,7)* vorab untersucht werden. Innerhalb des Forschungsprojektes sollen all diese Fragen geklärt sowie Methoden und Maßnahmen erarbeitet werden, um einen KI-Ansatz von Grund auf fair zu gestalten. Dabei müssen unterschiedliche Ansätze verfolgt werden sodass KI-Systeme

- im Nachhinein auf faire Vorhersagen kalibriert werden können,
- während des Trainings die Fairness in dessen Optimierungskriterium einbindet,
- nur mit Datenbasen trainiert werden die vorab bezüglich des Fairnessgedankens aufbereitet wurden.

Zudem müssen Metriken für die Bewertung der KI-Systeme definiert werden, sodass eine generelle Aussage über die Fairness eines KI-Systems getätigt werden kann.



Projektteam

Christopher Hadyk
christopher.hadyk@hs-ruhrwest.de
Prof. Dr.-Ing. Anselm Haselhoff
anselm.haselhoff@hs-ruhrwest.de
Prof. Dr.-Ing. Uwe Handmann
uwe.handmann@hs-ruhrwest.de

Gefördert durch das

Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen

im Rahmen des Förderprogramms FH Kompetenz

Hochschule Ruhr West, Institut Positive Computing, Lützowstraße 5, 46236 Bottrop, Germany