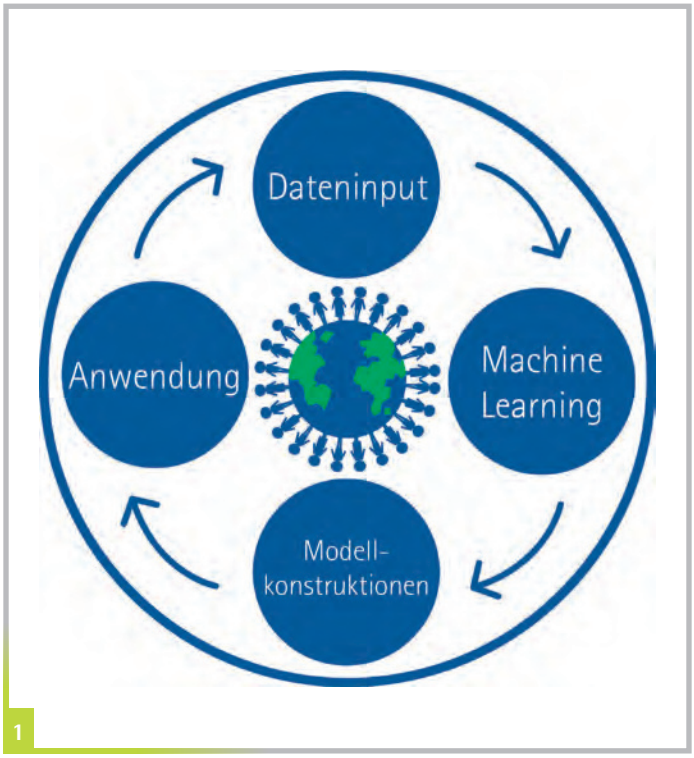
Big Data, Machine Learning

– und diskriminierende Algorithmen?

Computer-Algorithmen werden zunehmend eingesetzt, um menschliche Entscheidungen in wichtigen gesellschaftlichen Bereichen zu unterstützen, anzuleiten oder sogar zu ersetzen. Indessen häufen sich die Hinweise, dass Algorithmen mitunter ebenso voreingenommen und diskriminierend agieren können wie Menschen.

In dem interdisziplinären Forschungsprojekt „Bias and Discrimination in Big Data and Algorithmic Processing – BIAS" arbeiten Wissenschaftler*innen aus Philosophie, Rechtswissenschaft und Informatik daran, diskriminierende Algorithmen zu verstehen und zu verbessern.



1

Abbildung 1
Der Kreislauf von Daten, Machine Learning, Modellbildung und Anwendung
Grafik: LUH

Wer gibt vor, welche der über sieben Millionen Ergebnisse Ihrer Google-Suche mit dem Stichwort „Diskriminierung" ganz oben in der Liste erscheinen? Wer entscheidet, welche Buchempfehlungen Sie nach Ihrem amazon-Kauf der „Encyclopaedia of Biases and Heuristics" erhalten? Wer übernimmt die Vorprüfung Ihrer Kreditwürdigkeit, wenn Sie bei Ihrer Hausbank wegen eines Baudarlehens anfragen? Richtig: Computer-Algorithmen. Und warum meinen Sie, dass diese Algorithmen weniger vorurteilsbehaftet und tendenziös als Menschen agieren?

Böse Bilder

2015 meldeten empörte Nutzer, dass die Bilderkennungs-Software „Google Photos" Bilder von afroamerikanischen Personen in die Rubrik „Gorillas" einsortiert hatte. Die Nachricht erzeugte einen mittleren Aufruhr in den sozialen Medien, die Herstellerfirma entschuldigte sich umgehend bei den Betroffenen. Und ein weiteres Beispiel dafür, dass algorithmische Klassifikationen weder technisch einwandfrei noch sozial harmlos funktionieren müssen, war auf der Tagesordnung.

Die Ursache der Panne war nicht schwer zu identifizieren: Offenbar war die Software auf unzureichendem Bildmaterial trainiert worden. Zu wenige Fotos von afroamerikanischen Personen in den Datensets, von denen der Algorithmus lernte, hatten dazu geführt, dass ihm in der Anwendung peinliche Fehler unterliefen. „Garbage in, garbage out" – „Müll rein, Müll raus", nennen Computerwissenschaftler diesen Effekt nicht repräsentativen Trainingsmaterials. Falls jemand Zweifel hat: Wird der Algorithmus mit zu wenigen Fotos von weißhäutigen Menschen trainiert, stuft er sie als Hunde oder Seerobben ein.

Die Ursache für einen Fehler zu finden, heißt nicht, ihn sofort beheben zu können. Das richtige Trainieren von Algorithmen – aufgrund von Big Data und Machine Learning – ist eine aufwändige Kunst. Google selbst entschied sich seinerzeit, die Kategorie „Gorillas" schlichtweg aus der App zu streichen. Aber immerhin war der Fall insofern einfach zu behandeln, als eindeutig ein technisches Defizit vorlag und kein Zweifel an der Fehlleistung des Algorithmus herrschte.

Rechnen mit Wörtern

Andere Probleme „diskriminierender Algorithmen" erweisen sich als deutlich schwieriger zu erfassen und

zu lösen. Wie etwa soll man reagieren, wenn ein Algorithmus in seinen Vorhersagen diskriminierend zu agieren scheint, dabei aber letztlich nur bestehende Muster der sozialen Realität reproduziert?

„Word embeddings" sind Algorithmen, die Wörter der normalen Sprache zur weiteren Verarbeitung in Zahlentupel („Vektoren") übersetzen. Dabei versuchen sie, in diesen numerischen Darstellungen semantische Verbindungen festzuhalten, die sie im Datenmaterial entdecken: Sinnverwandte Wörter sollen mathematisch „nah" beieinanderliegen, Rechenoperationen sollen Bedeutungszusammenhänge wiedergeben. Lässt man solch einen Algorithmus aus üblichen Textkonvoluten lernen, wird er seine Zahlenkodierungen so einrichten, dass sinnvolle Gleichungen gelten, wie etwa: „King – Queen = Man – Woman". Aber er wird auch Relationen liefern wie: „Computer Programmer – Housekeeper = Man – Woman". – Problematisch?

Zumindest nicht überraschend, wenn man sich an das „garbage in, garbage out"-Prinzip erinnert. Nur hat man es diesmal nicht mit lückenhaftem Datenmaterial zu tun, sondern mit Daten, in denen sich jahrzehntealte geschlechterspezifische Stereotype unserer Gesellschaft abgebildet haben. Bislang sind mehr Männer Programmierer, und mehr Frauen erledigen Hausarbeit. Kann man es dem Algorithmus vorwerfen, dass er angesichts solcher Fakten nun diese Stereotype reproduziert? Dennoch scheint es eine Überlegung wert, seine Klassifikationen zu korrigieren, damit diese nicht beispielsweise bei einer Internet-Recherche weibliche Programmierer benachteiligen – und so ihrerseits dazu beitragen, dergleichen Zusammenhänge in unserer Gesellschaft zu verstetigen (*vgl. Abb. 1*).

Algorithmen putzen

Computerwissenschaftler und Computerwissenschaftlerinnen (nicht wenige!) arbeiten seit einiger Zeit intensiv an Prozeduren, um Algorithmen von dergleichen Einseitigkeiten zu befreien. „Debiasing" ist das Codewort, unter dem sich ihre Forschungsansätze versammeln. Die Herausforderung besteht darin, algorithmische Klassifikationen von ungewünschten Verzerrungen zu säubern, ohne dabei ihre Vorhersagekraft derart zu kompromittieren, dass sie nutzlos werden. Der Tradeoff zwischen „fairness" und „accuracy" ist zu einer wesentlichen Herausforderung automatisierter Datenanalyse geworden.

Aber diese Abwägung ist kein rein technisches Problem. Wer sich auf die harten Fälle der Debatte einlässt, merkt bald, dass klassische Fragen nach den sozialen Ursachen und dem angemessenen Umgang mit Einseitigkeiten und Diskriminierung in unserer Gesellschaft aufgeworfen werden, die Rechtswissenschaft und politische Philosophie schon seit Längerem beschäftigen, nun aber in ganz neuem Kontext erscheinen. Dies wird besonders deutlich, wenn es um Anwendungsfälle geht, die das Schicksal von Einzelnen und Gemeinschaften erheblich berühren.

Im Zweifel gegen den Angeklagten

Im Jahr 2016 publizierte die investigative Plattform ProPublica eine vieldiskutierte Studie mit dem Titel „Machine Bias" (https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing). Darin legten Julia Angwin et al. statistische Belege für die Behauptung vor, dass ein kommerzieller Computer-Algorithmus, der in einigen US-Staaten verwendet wird, um die Rückfallwahrscheinlichkeiten von Straftätern zu prognostizieren, Afroamerikaner systematisch benachteilige. Der zentrale Punkt ihrer Analyse bestand in der Feststellung, dass der Algorithmus „COMPAS" (= Correctional Offender Management Profiling for Alternative Sanctions) bei schwarzen Personen eine signifikant erhöhte Fehlerquote aufweise.

Genauer zeigte sich im Rückblick, dass COMPAS schwarze Delinquenten, die keine weitere Straftat begingen, mit einem Anteil von 45 Prozent fälschlich als „high risk" eingestuft hatte. Bei weißen Straffälligen lag diese Fehlerquote („false positive rate", FPR) demgegenüber nur bei 23 Prozent. Umgekehrt war die Fehleinstufung von späteren Wiederholungstätern als „low risk" („false negative rate", FNR) bei Weißen deutlich höher als bei Schwarzen, nämlich 48 Prozent gegenüber 28 Prozent. Die Betreiberfirma von COMPAS wies den Vorwurf der Diskriminierung indessen zurück mit dem Hinweis, dass die Quote richtiger Vorhersagen keine nennenswerte Differenz zwischen beiden Gruppen aufwies: Prüft man, wie viele der als „high risk" eingestuften Personen tatsächlich wieder straffällig wurden („positive predictive value", PPV), so liegt dieser Anteil bei 59 Prozent für weiße und bei 63 Prozent für schwarze Delinquenten (*s. Infokasten*).

Bei alledem ist zu beachten, dass COMPAS den Parameter Hautfarbe explizit überhaupt nicht benutzt. Die ungleichen Fehlerquoten für Weiße und Schwarze beruhen auf anderen Variablen wie Familienhintergrund, soziales Umfeld, Bildungsstand oder berufliche

Situation, die COMPAS über einen Fragebogen einholt und die in den USA stark mit der Hautfarbe korrelieren. Wenn man diese Variablen indessen herausnimmt, ist zu befürchten, dass auch Informationen verloren gehen, die für die Prognose wichtig sind.

Eine naheliegende Forderung wäre, dass all die genannten Maßzahlen (FPR, FNR, PPV …) für die beiden Gruppen (annähernd) gleich sein sollten. Leider ist dies jedoch unmöglich: Wenn zwei Gruppen unterschiedliche Prävalenzen bezüglich eines Merkmals aufweisen (und das ist in dem diskutierten Beispiel der Fall: Schwarze und Weiße haben in der Grundgesamtheit abweichende Grade von Rückfälligkeit), so ist es streng mathematisch ausgeschlossen, dass sämtliche Qualitätsmaße von Vorhersagen für diese beiden Gruppen gleich sind.

Eine zentrale Frage lautet also, welches Fairness-Maß im vorliegenden Fall vordringlich wäre: Sollte die Fehlerquote oder doch eher die Vorhersagestimmigkeit für die zwei Gruppen gleich sein – wenn man schon nicht beides haben kann? Wieder ist man zu Trade-off gezwungen. Und eine berechtigte Abwägung wird weder unabhängig von dem konkreten Anwendungsbereich (Verhängung von Haftstrafen) noch unabhängig von den historischen Hintergründen (insbesondere den Folgen von jahrhundertelanger Sklaverei und Unterdrückung der schwarzen Bevölkerung in den USA) zu treffen sein. Wer übrigens in diesem Beispiel einen Beweis dafür erkennen will, dass Algorithmen ohnehin nichts taugen, sollte sich nicht zu früh freuen: Die genannten Unmöglichkeitstheoreme, die dafür sorgen, dass sich unvermeidliche Ungleichheiten zwischen beiden Gruppen einstellen, hängen in keiner Weise davon ab, dass es sich um algorithmische Vorhersagen handelt. Würde man menschliche Entscheidungen auswerten, wären sie von genau demselben Dilemma betroffen.

Oder sollte man das Problem vielleicht völlig anders angehen? Sollte man statt statistischer Vergleiche lieber die individuelle Ebene heranziehen und untersuchen, ob COMPAS eine bestimmte Person anders eingestuft hätte, wenn sie eine andere Hautfarbe gehabt hätte?

Intuitiv ist dies ein überzeugender Zugang, um mögliche Diskriminierungen durch einen Algorithmus festzustellen. Aber wie genau soll man sich die Alternative vorstellen, dass Person X weiß statt schwarz wäre? Ändert sich nur ihre Hautfarbe? Nun, dann wird COMPAS keinen anderen risk score ausgeben, weil ihm die Hautfarbe der Personen überhaupt nicht vorliegt (*s.o.*). Oder sollte man realistischerweise annehmen, dass die Person dann auch eine andere Schulbildung, ein anderes Einkommen, ein anderes Wohnumfeld gehabt hätte? Wenn dies aber anzunehmen ist und COMPAS nun tatsächlich eine abweichende Prognose stellen würde – lässt sich dies noch als Diskriminierung auslegen, wo Parameter wie Schulbildung oder

Statistische Fairness: Der Fall COMPAS

	wurde wieder straffällig	wurde nicht wieder straffällig
als „high risk““ eingestuft	„true positive (TP)“ *w = 505; s = 1.369*	„false positive (FP)“ *w = 349; s = 805*
als „low risk“ eingestuft	„false negative (FN)“ w = 461; s = 532	„true negative (TN)“ w = 1.139; s = 990

Fehlermatrix für COMPAS: Anzahl der „true positives“, „false positives“, „false negatives“ und „true negatives“, jeweils für weiße (*w*) und schwarze (*s*) Delinquenten. Werte für Broward County, Florida (2013/14) (Zahlen übernommen aus www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm).

Statistische Qualitätsmaße:
„False positive rate (FPR)” = FP/{FP+TN}: Gegeben, man gehört zur Gruppe der nicht wieder straffällig werdenden Personen (FP+TN) – wie groß ist die Wahrscheinlichkeit, dass man fehleingestuft wird als „high risk“ (FP)? Ergebnis: FPRw = 23%, FPR_s = 45%.

„False negative rate (FNR)” = FN/{TP+FN}: Gegeben, man gehört zur Gruppe der wieder straffällig werdenden Personen (TP+FN) – wie groß ist die Wahrscheinlichkeit, dass man fehleingestuft wird als „low risk“ (FN)? Ergebnis: FNRw = 48%, FNR_s = 28%.

„Positive predictive value (PPV)“ = TP/{TP+FP}: Gegeben, man wurde als „high risk“ eingestuft (TP+FP) – wie groß ist die Wahrscheinlichkeit, dass man tatsächlich wieder straffällig wird (TP)? Ergebnis: PPV_w = 59%, PPV_s = 63%.

Statistische Fairnessmaße:
„Predictive equality“: gleiche FPR für beide Gruppen.
„Equal opportunity“: gleiche FNR für beide Gruppen.
„Equalized odds“: gleiche FPR und gleiche FNR für beide Gruppen.
„Predictive parity“: gleicher PPV für beide Gruppen.

Wohnumfeld wahrscheinlich prädiktiv für Straffälligkeit sind?

Schwarze Schachteln

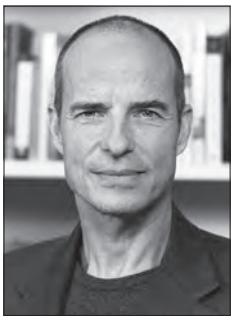
Neben diesen Fragen der statistischen Balance gibt es noch grundlegendere Aspekte, welche die Verwendung von Computerprognosen in Gerichtsprozessen problematisch machen. Algorithmen, die durch Machine Learning entwickelt werden, sind in der Regel „black boxes". Das heißt, man kennt ihre internen Prozesse nicht, und insbesondere ist nicht offensichtlich, auf welche Input-Parameter sie ihre Prognosen stützen. Im Rahmen einer Gerichtsverhandlung erscheint es indessen nicht unproblematisch, wenn Rückfallprognosen Haftentscheidungen begründen, ohne dass bekannt wäre, auf welche genauen Eigenschaften der angeklagten Person sich diese Entscheidungen letztlich stützen.

Dass die EU-Datenschutzrichtlinie im Falle von algorithmischen Entscheidungen von einem „right to explanation" spricht, ist vor diesem Hintergrund sehr gut nachvollziehbar. Aber wie weit dieses Recht gehen sollte und was es konkret impliziert, wird sorgfältig in verschiedenen Anwendungsbereichen zu prüfen sein.

Man sieht: Was als ein technisches Problem der Computerwissenschaften begann, weist unversehens weiter in komplexe Bereiche von Rechtswissenschaft und Philosophie, berührt unsere fundamentalen Vorstellungen von Fairness und wirft alte Diskussionen um gesellschaftlichen Ausgleich, den Umgang mit historischem Unrecht und politische Maßnahmen wie Affirmative Action auf.

Ob wir diese Herausforderungen bestehen werden, bleibt abzuwarten. Die Algorithmen werden indessen nicht warten. Und das Thema Diskriminierung zeigt, dass das Science-Fiction-Szenario einer KI, die menschliche Intelligenz übersteigt, vielleicht nicht das einzige Problem ist, vor das Algorithmen uns stellen könnten: Eine KI, die unsere Fehler einfach nachmacht, wird uns ebenso beschäftigen. Und sie ist bereits unter uns.

Prof Dr. Dietmar Hübner
Jahrgang 1968, ist Professor für praktische Philosophie, insbesondere Ethik der Wissenschaften, am Institut für Philosophie an der Philosophischen Fakultät. Seine Forschungsschwerpunkte liegen in der allgemeinen Ethik, der angewandten Ethik, der politischen Philosophie sowie im Themenkreis Willensfreiheit und Verantwortlichkeit. Kontakt: dietmar.huebner@philos.uni-hannover.de



Prof. Dr. Mathias Frisch
Jahrgang 1964, ist Professor für theoretische Philosophie, insbesondere Wissenschaftsphilosophie, am Institut für Philosophie an der Philosophischen Fakultät. Seine Arbeitsschwerpunkte sind Philosophie der Klimawissenschaften, Philosophie der Physik, sowie allgemeine Wissenschaftsphilosophie. Kontakt: mathias.frisch@philos.uni-hannover.de

Prof. Dr. Uljana Feest
Jahrgang 1967, ist Professorin für Philosophie der Sozialwissenschaften und Sozialphilosophie am Institut für Philosophie der Leibniz Universität. Ihre Arbeitsschwerpunkte liegen in den Bereichen der Philosophie und Geschichte der Psychologie sowie der Philosophie des Experimentes. Sie beschäftigt sich u.a. mit Fragen rund um die Konzeptualisierung und Erforschung impliziter Biases in der Psychologie. Kontakt: feest@philos.uni-hannover.de