Fairness und Qualität algorithmischer Entscheidungen

Zweig, Katharina A.; Krafft, Tobias D.

Veröffentlichungsversion / Published Version
Sammelwerksbeitrag / collection article

Empfohlene Zitierung / Suggested Citation:
Zweig, K. A., & Krafft, T. D. (2018). Fairness und Qualität algorithmischer Entscheidungen. In R. Mohabbat Kar, B. E. P. Thapa, & P. Parycek (Hrsg.), *(Un)berechenbar? Algorithmen und Automatisierung in Staat und Gesellschaft* (S. 204-227). Berlin: Fraunhofer-Institut für Offene Kommunikationssysteme FOKUS, Kompetenzzentrum Öffentliche IT (ÖFIT). https://nbn-resolving.org/urn:nbn:de:0168-ssoar-57570-1

Fairness und Qualität algorithmischer Entscheidungen

Katharina A. Zweig & Tobias D. Krafft

Technische Universität Kaiserslautern

Algorithmische Entscheidungssysteme werden immer häufiger zur Klassifikation und Prognose von menschlichem Verhalten herangezogen. Hierbei gibt es einen breiten Diskurs um die Messung der Entscheidungsqualität solcher Systeme (Qualität) und die mögliche Diskriminierung von Teilgruppen (Fairness), welchen sich dieser Artikel widmet. Wir zeigen auf, dass es miteinander unvereinbare Fairnessmaße gibt, wobei wir auf zwei im Speziellen eingehen. Für sich allein betrachtet sind die zwei Maße zwar logisch und haben je nach Anwendungsgebiet auch ihre Daseinsberechtigung, jedoch können nicht beide zugleich erfüllt werden. Somit zeigt sich, dass gerade im Einsatz algorithmischer Entscheidungssysteme im Bereich der öffentlichen IT aufgrund ihres großen Wirkungsbereichs auf das Gemeinwohl höchste Vorsicht bei der Wahl solcher Maßstäbe herrschen muss. Wird im Anwendungsfall die Erfüllung sich widersprechender Maßstäbe gefordert, so muss darüber nachgedacht werden, ob eine algorithmische Lösung an dieser Stelle überhaupt eingesetzt werden darf.

1. Einleitung

Menschen müssen ständig andere Menschen bewerten: von Leistungsbewertungen in Ausbildung und Beruf über die Kreditwürdigkeit oder möglicherweise zu berücksichtigende mildernde Umstände

bei Gerichtsurteilen ist der Mensch dem Urteil anderer Menschen ausgesetzt. Die letzten Jahrzehnte sind allerdings geprägt von einer psychologischen Forschung, die verschiedene Unzulänglichkeiten von menschlichen Entscheiderinnen und Entscheidern zu Tage gebracht haben, darunter insbesondere die Forschungen von Kahnemann und Tversky[1] und von Ariely.[2]

Im juristischen Bereich konnte beispielweise gezeigt werden, dass die Risikofreudigkeit bei Richterinnen und Richtern bei Anträgen auf vorzeitige Haftentlassung mit Fortschreiten der Tagesstunde abnahm - um nach dem Mittagessen wieder zu steigen.[3] Auch andere Indizien weisen darauf hin, dass Richterinnen und Richter nicht immer ganz vorurteilsfrei entscheiden: In den USA findet sich der global zweithöchste Anteil an Staatsbürgern im Gefängnis (666 von 10.000 Einwohnern),[4] und davon sind weit überproportional viele Afroamerikaner betroffen.[5] Es wird geschätzt, dass jeder dritte afroamerikanische Junge in den USA in seinem Leben mindestens einmal im Gefängnis sitzen wird.[6] Unter diesen Umständen rufen auch Bürgerrechtsbewegungen wie die American Civil Liberty Union (ACLU) dazu auf, die Entscheidungsprozesse zu objektivieren und transparenter zu gestalten und weniger dem vermeintlich unzuverlässigen und irrationalen Handeln anderer Menschen zu überlassen.[7] Dabei liegt die Hoffnung auf Algorithmen des maschinellen Lernens und der künstlichen Intelligenz, die z. B. eine Rückfällig-

1 Kahnemann, 2012
2 Ariely, 2010
3 Danziger, 2015
4 Statista, 2018
5 ACLU, 2011
6 The Sentencing Project, 2013
7 ACLU, 2011

keitsvorhersage bei schon vorher straffällig gewordenen treffen sollen. Solche Algorithmen werden momentan in einigen Bundesstaaten schon nach einer Haftentlassung verwendet, z. B. um seltene Therapieplätze an diejenigen mit dem höchsten Risiko zu verteilen, und in manchen Bundesstaaten auch während des Prozesses und vor dem eigentlichen Urteilsspruch.[8]

Diese Art der ›algorithmischen Entscheidungssysteme‹ sind zunehmend häufig anzutreffen und bedürfen insbesondere dann, wenn die Systeme direkt oder indirekt über das Leben von Menschen entscheiden und damit deren gesellschaftliche Teilhabe vergrößern oder auch verkleinern können, der gesellschaftlichen Qualitätskontrolle.[9] Im Folgenden werden wir zuerst einen Überblick geben, wie algorithmische Entscheidungssysteme - basierend auf Daten - Entscheidungsregeln extrahieren, die dann auf weitere Personen angewendet werden können (Abschnitt 2). In der Informatik werden die Entscheidungen dieser Systeme im Wesentlichen mit Hilfe eines einzigen Maßes auf ihre Qualität hin evaluiert - dazu wird vom jeweiligen Designteam eines von mehreren Dutzend möglicher Qualitätsmaße ausgewählt. Manche Systeme werden zusätzlich noch mit sogenannten »Fairnessmaßen« daraufhin analysiert, ob sie unzulässige Diskriminierungen vornehmen. Auf die Auswahl und die Bedeutung dieser Fairnessmaße konzentrieren wir uns in diesem Artikel, während unser zweiter Artikel in diesem Band die Wahl des Qualitätsmaßes diskutiert.[10] In Abschnitt 3 gehen wir auf der einen Seite darauf ein, welche Methoden es für die Messung der Entscheidungsqualität solcher Systeme gibt, und auf der anderen Seite darauf, wie eine mögliche Diskriminierung von Teilgruppen der Bevölkerung

[8] EPIC, 2017

[9] Lischka & Klingel, 2016

[10] Krafft & Zweig, 2018

evaluiert werden kann. Es stellt sich heraus, dass es für beide Aspekte eine Vielzahl von Maßen gibt und es daher nicht immer eindeutig ist, welches davon eingesetzt werden muss. Wir zeigen weiterhin auf, dass dies noch nicht einmal von der Art des Entscheidungssystems an sich abhängt, sondern abhängig vom Kontext des Einsatzes des Systems gewählt werden muss, also dem sozialen Prozess, in dem die Entscheidung getroffen wird. In Abschnitt 4 fassen wir daher zusammen, welche Designentscheidungen bei der Entwicklung eines algorithmischen Entscheidungssystems und bei seiner Einbettung in ein soziales System getroffen werden müssen und warum wir einen interdisziplinären und gesellschaftlichen Diskurs über diese relativ technischen Fragen benötigen.

2. Algorithmische Entscheidungssysteme

Algorithmen sind eindeutig definierte Lösungswege, um für ein festgelegtes Problem eine korrekte Lösung zu finden.[11] Ein Beispiel aus dem Alltag stellt das Navigationsproblem dar, das für zwei Orte danach fragt, wie man am schnellsten von A nach B kommt. Für dieses Problem gibt es mehrere Algorithmen, also mehrere Verfahren, die schnellste Route zu identifizieren - sie alle haben aber dasselbe Ergebnis, kommen nur auf unterschiedlichem Weg dorthin. Die Algorithmenentwicklung hat im Lauf der letzten Jahrzehnte Tausende von klassischen Problemen beschrieben, die jeweils von einer Vielzahl an Algorithmen gelöst werden können. Darunter sind einige, für die die Berechnung schnell geht, und andere, für die es nahezu unmöglich ist, die wirklich beste Lösung in angemessener Zeit zu berechnen.[12]

[11] Zweig, 2016a

[12] Stiller, 2015

»Algorithmische Entscheidungssysteme« sollen Menschen dabei unterstützen, Dinge oder Personen zu klassifizieren. Ganz allgemein definieren wir ein algorithmisches Entscheidungssystem als eine Software, die aus einer ganzen Reihe von Eingabewerten einen einzigen Wert berechnet. Dieser wird entweder als eine Klassifikation interpretiert oder bewertet das dazugehörige Objekt bzw. Subjekt auf einer Skala (siehe Abbildung 1). Erfolgreiche algorithmische Entscheidungssysteme werden z. B. in der Produktion verwendet, um automatisch fehlerhafte Bauteile zu identifizieren[13] - sie sind häufig sogar vollautomatisch, steuern also direkt Gebläse oder Roboterarme an, welche die als fehlerhaft erkannten Dinge aus dem Produktionsprozess entfernen. Als »Eingabe« bekommen diese Systeme beispielsweise Kamerabilder, auf denen sie nach Hinweisen darauf suchen, dass das Produkt fehlerhaft ist. Im Allgemeinen sind algorithmische Entscheidungssysteme sehr erfolgreich und bilden die Grundlage für Internetsuchmaschinen, Produktempfehlungssysteme, autonomes Fahren und flexibel einzusetzende Roboter.

Wenn es allerdings darum geht, Personen zu klassifizieren oder ihr zukünftiges Verhalten basierend auf Analogien zu bisherigem Verhalten anderer Personen vorherzusagen (siehe Abbildung 1), kann es dazu kommen, dass die gesellschaftliche Teilhabe der algorithmisch bewerteten Personen behindert wird.[14] Klassische Anwendungsgebiete solcher algorithmischer Entscheidungssysteme sind die Bewertung der Kreditwürdigkeit von Personen im Sinne einer Vorhersage des Ausfallrisikos[15] (z.B. Schufa) oder das sogenannte *»People Analytics«* als die Anwendung von Data-Science-Methoden,

[13] Schramm & Schramm, 1990

[14] Lischka & Klingel, 2017

[15] Robinson, David & Harlan Yu & 2014

um die Leistung von Personen im Arbeitsleben zu bewerten.[16] In diesem Artikel nehmen wir nur solche algorithmischen Entscheidungssysteme in den Fokus, die tatsächlich Menschen bewerten, und zwar, indem sie eine Reihe von Eigenschaften der Personen als Eingabe bekommen und als Ergebnis einen einzigen Wert berechnen. Ein Beispiel dafür wäre die Einteilung von Personen in unterschiedliche Schadenfreiheitsklassen aufgrund ihrer bisherigen Fahrerfahrung oder die Einordnung von Personen in Gefährderklassen bezüglich einer möglichen terroristischen Aktivität.[17] Oftmals beruht die finale Ausgabe des Systems in einem Zwischenschritt auf einer numerischen Bewertung, die die Personen erst einmal relativ zueinander ordnet. Diese ermöglicht also zu sagen: »Person A ist eher eine Terroristin als Person B«. Eine nachgelagerte Schwellwertbestimmung erlaubt es dann, die »vermeintlichen Terroristen« in die eine Kategorie einzusortieren und die »vermutlich normalen Bürger« in die andere Klasse einzuordnen. Solche Bewerter, die eine Einsortierung berechnen, werden *»Scoring«*-Funktionen genannt. Darauf basierend kann dann ein relatives Ranking berechnet werden oder der Wert wird als absolute Zahl verwendet, um eine Entscheidung zu treffen. Ein Beispiel für Letzteres ist das algorithmische Entscheidungssystem hinter dem *»China Citizen Score«,* mit dem das bürgerliche Verhalten aller chinesischen Bürger bewertet werden soll und von dem ebenfalls Kredit- und Visa-Vergabe-Entscheidungen abhängen sollen.[18]

Eine algorithmische Klassifikation von Personen kann auch als Grundlage für eine Vorhersage genutzt werden (s. Abbildung 1c). Das algorithmische Entscheidungssystem bekommt dazu die Eigenschaften einer Person als Grundlage und entscheidet, welche Gruppe

[16] Reindl & Krügl, 2017

[17] The Intercept, 2015

[18] Helbing et. al., 2017; Assheuer, 2017

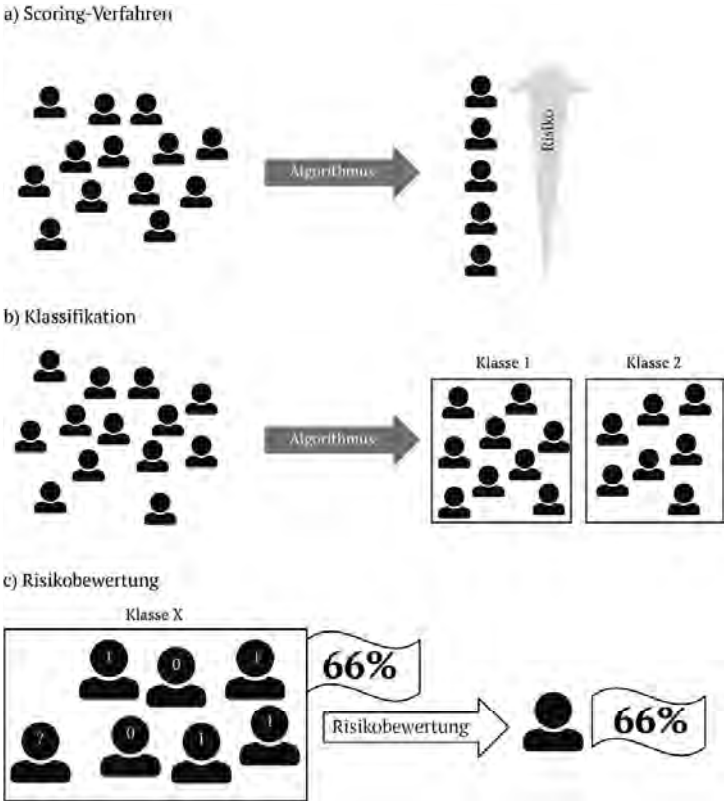

Abbildung 1: Algorithmische Entscheidungssysteme, die Menschen bewerten, weisen diesen entweder einen »Score« zu (a) oder teilen diese in Klassen (b) ein. Dementsprechend spricht man von Scoring-Verfahren und von Klassifikatoren. Mit einer Klassifikation kann dann die Rate, mit der das vorherzusagende Verhalten in der zugeordneten Klasse bisher auftrat (symbolisiert durch eine 1), der zu bewertenden Person als persönliches Risiko zugewiesen werden (c). Ein bekanntes - wenn auch simples - Beispiel für ein Scoring-Verfahren ist die Schufa, die einen Kreditwürdigkeits-Score berechnet. Ein Beispiel für einen ebenfalls sehr einfachen Klassifikator ist die Einteilung von Autofahrerinnen und Autofahrern in verschiedene Schadensfreiheitsklassen. In diesem Artikel geht es um solche algorithmischen Entscheidungssysteme, bei denen die Entscheidungsregeln aus Daten mit Hilfe von maschinellem Lernen abgeleitet wurden.

von Personen mit bekanntem Verhalten am ähnlichsten zu dieser Person ist. Der Anteil der Personen in dieser Gruppe, die das gesuchte Verhalten aufweisen, wird dann als Wahrscheinlichkeit interpretiert, dass die Person das Verhalten in der Zukunft zeigen wird. Auch dazu ein Beispiel: Wenn sich herausstellt, dass von den meisten der festangestellten Kreditbewerberinnen mit einem monatlichen Gehalt von mindestens 3.000 € ein Kredit in Höhe von 200.000 € zurückgezahlt wird, sieht es für Antragssteller mit denselben Eigenschaften gut aus.

Es gibt viele Methoden, Personen so in Gruppen aufzuteilen, dass möglichst viele Personen innerhalb derselben Gruppe dasselbe Verhalten zeigen. In den meisten Situationen gibt es keine perfekte Aufteilung. Daher müssen die vom System getroffenen Entscheidungen in ihrer Qualität durch sogenannte »Qualitätsmaße« bewertet werden, um zu erkennen, ob das System schon hinreichend gut ist, um in der Praxis genutzt zu werden. Diese Qualitätsmaße betrachten dabei immer die Qualität über alle Entscheidungen hinweg, während »Fairnessmaße« darüber hinaus bewerten, ob dieselbe Qualität auch für durch das Recht geschützte Teilgruppen gilt, ob also die Entscheidungsqualität nicht zwischen Teilgruppen diskriminiert. Im Folgenden werden diese Maße genauer diskutiert.

2.1. Qualität und Fairness eines algorithmischen Entscheidungssystems

Die vom System vorgenommenen Einteilungen können hinsichtlich ihrer Qualität und Fairness auf verschiedene Arten und Weisen bewertet werden. Alle Methoden beruhen darauf, dass man das System Entscheidungen auf einem Datensatz treffen lässt, bei dem die tatsächliche Klassifizierung bekannt ist. Im Falle der Kreditwürdigkeit wäre dies ein Datensatz, bei dem die Personen ihren Kredit schon zurückgezahlt haben bzw. versäumt haben, diesen zurückzuzahlen.

Soll ein algorithmisches Entscheidungssystem terroristische Aktivitäten identifizieren, wird der Algorithmus auf Personen getestet, die nachweislich für terroristische Aktivitäten verurteilt wurden. In beiden Fällen wird die Klasse, die man gerne identifizieren will, als die »positive« Klasse bezeichnet. Hier handelt es sich nicht um ein Werturteil, die Benennung stammt aus der medizinischen Praxis. Wenn ein Test auf eine Krankheit »positiv« ausfällt, ist dies auch kein Werturteil, sondern zeichnet dasjenige Ergebnis aus, nachdem gesucht wird: In diesem Sinne kann man das Auffinden von Kreditwürdigen und das Auffinden von Terroristen als den »positiven« Fall bezeichnen, ohne dies inhaltlich zu bewerten. In beiden Situationen gibt es vier Fälle, die eintreten können:

1. Kreditwürdige werden als kreditwürdig erkannt und Terroristen als Terroristen. Der Algorithmus hat die zu identifizierenden Fälle erkannt. Wir sprechen von *›true positives‹*.

2. Nicht-kreditwürdige bzw. unschuldige Bürger werden als nicht-kreditwürdig bzw. nicht-terroristisch erkannt. Wir sprechen von *›true negatives‹*.

3. Wenn Nichtkreditwürdige fälschlich als kreditwürdig und unschuldige Bürger fälschlich als Terroristen bezeichnet werden, sprechen wir von *›false positives‹* (falsch positive Entscheidungen). Diese Fälle verursachen einen hohen institutionellen bzw. individuellen Schaden, sind aber nicht völlig zu vermeiden, wenn Banken Gewinnchancen maximieren wollen und die Gesellschaft vor möglichst allen Terroristen und Terroristinnen geschützt werden soll.

4. Wenn Kreditwürdige oder Terroristen als nicht kreditwürdig bzw. unschuldige Bürger bezeichnet werden, sprechen wir von *›false negatives‹* (falsch negative Entscheidungen). Bei diesen

Fällen liegt der Schaden beim Individuum bzw. der Gesamtgesellschaft.

Wie beschrieben haben also die Fehlurteile individuelle und gesellschaftliche Kosten und unterschiedliche Institutionen haben verschiedene Ansichten darüber, wie diese zu gewichten seien: Wiegt der fälschlich als Terrorist bezeichnete Bürger schwerer als der nicht entdeckte Terrorist? Müssen Banken gesellschaftliche Chancen an viele verteilen oder sollten sie möglichst wirtschaftlich arbeiten? Zu diesen und anderen Fragen haben naturgemäß Verbraucherschutzorganisationen, Bürgerrechtsorganisationen und Geheimdienste höchst unterschiedliche Ansichten, die unterschiedliche Gewichtungen der Anzahlen an korrekten bzw. falschen Entscheidungen verlangen. Neben diesen einfachen, gewichteten Summen der Anzahlen korrekter und falscher Entscheidungen gibt es aber auch anders geartete Qualitätsmaße, deren Verwendung vom sozialen Prozess abhängt.[19]

Hier wenden wir uns der Frage zu, inwieweit die - möglicherweise falsche - Einteilung der Personen in die verschiedenen Klassen sich auf schützenswerte Minderheiten oder Teilgruppen auswirkt. Dies lässt sich am besten am Beispiel der Rückfälligkeitsvorhersage demonstrieren, wo es gilt, die Klasse der vermutlich rückfällig Werdenden zu identifizieren - dies ist die »positive Klasse« (wieder ohne Werturteil). Es klingt zuerst völlig selbstverständlich, dass z. B. Frauen genauso oft »falsch positiv« in Hochrisikogruppen eingeteilt werden sollten wie Männer. Im Allgemeinen würde man vielleicht fordern, dass alle vom Antidiskriminierungsverbot genannten Gruppen in jeweils demselben Anteil in den »falsch positiven« zu finden sein sollten wie sie auch in der Bevölkerung vorkommen. Diesen Aspekt von Diskriminierungsfreiheit untersuchte die journalistische

[19] siehe dazu Krafft & Zweig, 2018

Plattform ProPublica und ihr Ergebnis hat für viel Aufmerksamkeit gesorgt:[20] Julia Angwin und ihre Kollegen fanden heraus, dass Afroamerikaner weit häufiger fälschlich in die Hochrisiko-Klasse eingeordnet werden, als sie tatsächlich rückfällig werden, und Weiße zu wenig häufig, gemessen an ihrer tatsächlichen Rückfälligkeitsquote. ProPublica nannte das algorithmische Entscheidungssystem COMPAS daraufhin ›rassistisch‹ und ›unfair‹. Ihr »Fairness«-Maß quantifiziert also, wer die Last der *›false positives‹,* der fälschlicherweise in die Hochrisikoklasse Einkategorisierten, zu tragen hat.

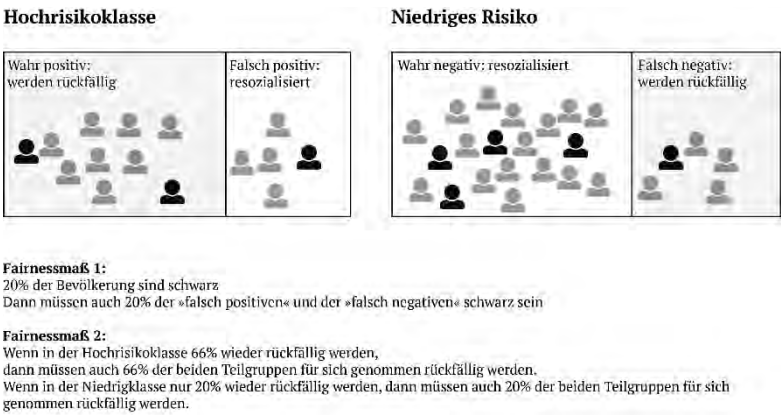

Abbildung 2: Zwei verschiedene Fairnessmaße am Beispiel von Rückfälligkeitsvorhersagen. In diesem Fall sind sie miteinander kompatibel, da beide Bevölkerungsgruppen, die »grauen« und die »schwarzen«, dieselbe Rückfälligkeitsquote haben: Von 32 »Grauen« werden 12 rückfällig, von 8 »Schwarzen« werden 3 rückfällig. Weichen die Rückfälligkeitsquoten der beiden Bevölkerungsgruppen stark voneinander ab, können nicht beide Fairnessmaße gleichzeitig erreicht werden.[21]

[20] Angwin, Larson, Mattu & Kirchner, 2017

[21] Kleinberg, Mullainathan & Raghavan, 2016

Die Firma *equivant* (vormals Northpointe Inc.), die COMPAS entwickelt und vertreibt, wehrte sich mit deutlichen Worten:[22] Dies sei nicht das in der Community übliche Fairnessmaß. Stattdessen müsste man das System daran messen, ob dieselbe Klassifizierung auch dasselbe für alle relevanten Teilklassen bedeute. Dazu muss man wissen, dass COMPAS Kriminelle zuerst in zehn Risikoklassen einteilt, die dann in drei Gruppen zusammengefasst werden. Für *equivant* ist das folgende Fairnessmaß das ausschlaggebende: Wenn Personen verschiedener Teilgruppen in Klasse 8 kategorisiert werden und die Klasse insgesamt eine Rückfälligkeitsquote von 60 Prozent aufweist, muss das auch für jede der diskriminierungsrelevanten Teilgruppen gelten. Dasselbe »Vor-Urteil« der Maschine muss also für Mann oder Frau, Afroamerikaner oder Weißen dasselbe bedeuten. Gemessen an diesem zweiten Fairnessmaß ist der COMPAS-Algorithmus tatsächlich »fair« und diskriminiert nicht.

Das ist aber noch nicht das Ende der Geschichte: Der Informatiker Jon Kleinberg fand zusammen mit Kollegen heraus, dass sich die Gesellschaft entscheiden muss, welches dieser beiden Fairnessmaße zum Zuge kommen soll, da sie in den meisten Fällen nicht kompatibel sind, nämlich dann, wenn die untersuchte Eigenschaft in den Teilgruppen unterschiedlich oft vorkommt.[23] Das ist in der Rückfälligkeitsquote tatsächlich der Fall: Frauen werden weniger oft rückfällig als Männer und Weiße weniger oft als Afroamerikaner. Dabei ist es unerheblich, aus welchen Gründen die Rückfälligkeitsquoten unterschiedlich sind - daher steht diese Frage auch nicht im Fokus dieses Artikels. Der folgende Abschnitt enthält eine Analogie, die erklärt, warum und wann die beiden Fairnessmaße nicht miteinander kompatibel sind. Er kann übersprungen werden, ohne dass das Textverständnis verloren geht.

[22] Dietrich, Mendoza & Brennan, 2016

[23] Kleinberg, Mullainathan & Raghavan, 2016

2.2. Warum die beiden Fairnessmaße miteinander nicht kompatibel sind

Warum unterschiedliche Rückfälligkeitsquoten problematisch sind, wenn man erreichen möchte, dass dieselbe Einordnung auch dasselbe bedeutet, wird an Abbildung 3 deutlich. Hier werden die Schwarzen öfter rückfällig als die Grauen (45 von insgesamt 105 Schwarzen, ca. 43 Prozent). Bei den Grauen werden auch 45 rückfällig, aber von insgesamt 135 Personen - eine Rückfälligkeitsquote von nur 33 Prozent. Der Algorithmus soll nun in zwei Klassen einteilen, in der sowohl die Schwarzen als auch die Grauen die jeweils gleichen Rückfälligkeitsquoten haben, z. B. bei ca. 84 Prozent in der Hochrisikoklasse und ca. 22-23 Prozent in der Niedrigrisikoklasse.[24] Dies ist also die Forderung von Fairnessmaß 2, die hier erfüllt werden soll und auch erfüllt werden kann.

Wir wollen nun zeigen, dass alleine diese Forderung – egal, wie gut sie umgesetzt wird – dazu führt, dass von derjenigen Gruppe, die die höhere Anfangsquote hat, mehr falsch-Positive in die Hochrisikoklasse kommen müssen. Um diese Verhältnisse zu schaffen, müssen in der Hochrisikogruppe ca. jeweils 5 Rückfällige auf 1 Resozialisierten kommen. Man kann also - in einem Gedankenexperiment - immer Kleingruppen von 6 Personen in dieser Art und Weise aus der jeweiligen Bevölkerungsgruppe nehmen. Aber nur, bis der Anteil von Rückfälligen im Rest eben ca. 22-23 Prozent Prozent beträgt.

Am Anfang sind also gar keine Schwarzen in der Hochrisikoklasse, dann 5 Rückfällige und 1 Resozialisierter, dann 10 und 2, 15 und 3

[24] Die Werte müssen sich nicht auf 100 Prozent ergänzen, weil sie sich jeweils auf die Personen in derselben Klasse beziehen. Da es um einzelne Menschen geht, können die gewünschten Prozentzahlen auch selten ganz genau erreicht werden.

und so weiter. Dabei bleibt das Verhältnis von rückfälligen zu resozialisierten Schwarzen immer gleich. In der verbliebenen, nicht Hochrisikoklasse verändert sich der Anteil der Rückfälligen von 45÷(45 Rückfällige + 60 Resozialisierte) = 43 % zu Beginn über 40÷(40+59) = 40 % nach dem ersten Schritt zu 35÷(35+58) = 37,7 % und nach dem zweiten Schritt und letztendlich nach dem sechsten Schritt zu 15÷(15+54) = 21,7 %.

Bei den Grauen müssen wir schneller stoppen, denn sie starten schon niedriger: Nämlich mit einer Quote an Rückfälligen von nur 45÷(45+90) = 33,3 %. Nach dem ersten Schritt sind noch 40÷(40+89) = 31 %, nach dem zweiten Schritt noch 35÷(35+88) = 28,5 %, und nach weiteren zwei Schritten nur noch 25÷(25+86) = 22,5 % rückfällig.

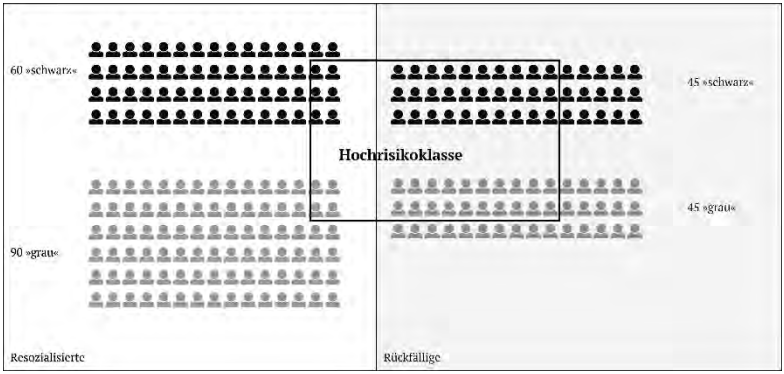

Abbildung 3: Wenn unterschiedliche Rückfälligkeitsquoten vorliegen, ist es nicht mehr möglich, beide Fairnessmaße (siehe Abbildung 2) gleichzeitig zu erfüllen. Die Hochrisikoklasse wird hier durch den schwarzen Kasten repräsentiert: Hier werden die schwarzen und die grauen Personen jeweils zu ca. 84 Prozent rückfällig (30 von 36 Schwarzen und 20 von 24 Grauen). Die übrig gebliebenen Grauen und Schwarzen werden jeweils zu ca. 22-23 Prozent rückfällig (15 von 69 Schwarzen und 25 von 111 Grauen).

Wir haben also denselben Schritt 6-mal bei den Schwarzen gemacht und nur 4-mal bei den Grauen. Damit ist offensichtlich, dass von den insgesamt 10 Personen, die falsch positiv in die Hochrisikoklasse gesteckt wurden, 60 Prozent schwarz sind, obwohl die Schwarzen insgesamt nur 105 von 240 Personen ausmachen, also deutlich weniger als die Hälfte der Bevölkerung. Dies Problem taucht immer auf, da die Rückfälligen und die Resozialisierten in demselben Verhältnis aus der Menge der »Schwarzen« und der »Grauen« entnommen werden.

Das heißt, dass durch die erhöhte Anfangsquote in einer der beiden Teilgruppen und die Forderung nach Fairnessmaß 2, es *immer* dazu kommt, dass die Last der falsch-positiven Entscheidungen in jener Teilgruppe höher ist, wo die Anfangsquote höher ist. Dies kann so extrem werden, dass alle falsch-positiven Entscheidungen in einer der beiden Teilgruppen liegen. In diesem Sinne kommt es unter diesem von der Firma *equivant* verwendeten Fairnessmaß unweigerlich dazu, dass die Personen der Gruppe mit der höheren Rückfälligkeitsquote auch prozentual häufiger in den Hochrisikoklassen sind und dann dort auch einen höheren Anteil von falsch-positiven Entscheidungen ausmachen.

Wie oben schon erwähnt, sind also auf der einen Seite Qualitätsmaße von vielen gesellschaftlichen Entscheidungen abhängig und Fairnessmaße können sich widersprechen. Die hier skizzierten Probleme bei der Wahl des richtigen Qualitätsmaßes und Fairnessmaßes bzw. der richtigen Maße zeigen damit, dass die dafür notwendigen Diskussionen nicht nur von einem kleinen Team an Informatikern geführt werden dürfen, sondern eines breiter geführten Diskurses bedürfen.

3. Gesellschaftlich notwendige Diskurse zur Evaluation von Qualität und Fairness gesellschaftlich relevanter algorithmischer Entscheidungssysteme

Um qualitativ hochwertige algorithmische Entscheidungssysteme zu konstruieren, müssen sowohl allgemeine Diskurse geführt werden, die den Gesamtprozess einer qualitätsgesicherten Konstruktion dieser Systeme festlegen, als auch jeweils spezifische Abwägungen durchgeführt werden, um den bestmöglichen und gerechten Einsatz eines solchen Systems in einem konkreten Kontext zu begleiten.

3.1. Allgemeine Diskurse zur Qualitätssicherung in der Konstruktion und in der Verwendung algorithmischer Entscheidungssysteme

Einen der wichtigsten gesellschaftlichen Diskurse stellt die Frage dar, ob es verallgemeinerte Situationen gibt, in denen algorithmische Entscheidungssysteme gar nicht eingesetzt werden sollten. Es zeichnet sich ab, dass es diese Situationen gibt (siehe Harcourt,[25] der gegen ihren Einsatz im Rechtssystem argumentiert). Die zweite zentrale Frage ist, nach welchen Dimensionen die potenzielle Schadenstiefe von Fehlentscheidungen durch algorithmische Entscheidungssysteme kategorisiert werden sollte, da die mögliche Schadenstiefe wiederum die Eingriffstiefe in den Konstruktionsprozess eines algorithmischen Entscheidungssystems rechtfertigt. Auch hier zeichnen sich erste Lösungen ab. Beispielsweise kann unterschieden werden zwischen solchen Systemen, welche die gesellschaftliche Teilhabe Einzelner stark einschränken könnten,[26] und

[25] Harcourt, 2006

[26] Lischka & Klingel, 2017

solchen, die Öffentlichkeit erzeugen und damit auf dieser Ebene schaden können.[27] Zur ersteren Kategorie gehören beispielsweise algorithmische Entscheidungssysteme, die Terroristen identifizieren sollen. Zur zweiteren gehören algorithmische Entscheidungssysteme auf sozialen Netzwerken, bei denen der individuelle Schaden eher klein ist, wenn relevante Nachrichten nicht angezeigt werden. In der Gesamtheit aller Nutzer können aber - wenn es sich um nicht angezeigte, aber politisch relevante Nachrichten handelt - Gesellschaften darunter leiden, dass Nachrichten politisch selektiert werden.[28]

Eine andere Dimension wird sein, inwieweit ein Markt existiert, der es erlaubt, dass sich Personen eine zweite Meinung einholen: Je monopolistischer ein algorithmisches Entscheidungssystem verwendet wird, desto besser muss die Qualitätssicherung sein. Dies gilt z. B. auch, wenn es in Deutschland zwar mehrere algorithmische Entscheidungssysteme gibt, die für eine automatische Leistungsbewertung von Mitarbeitern genutzt werden, die eigene Firma aber nur eines davon verwendet. Andere algorithmische Entscheidungssysteme werden einer Versicherungspflicht unterliegen, beispielsweise im Bereich des autonomen Fahrens. Auch das wird einen Anreiz für die Entwicklung bessere Entscheidungssysteme schaffen, die dann aber immer noch hinsichtlich weiterer Qualitätsaspekte bewertet werden müssten.

Im Rahmen der zunehmenden Nutzung algorithmischer Entscheidungssysteme - insbesondere solcher, die über die gesellschaftliche Teilhabe von Menschen direkt oder indirekt mitentscheiden[29] - müssen wir uns als Gesellschaft auch neu mit dem Begriff des Vorurteils

[27] Jaume-Palasí & Spielkamp, 2017

[28] Lischka & Stöcker, 2017

[29] Lischka & Klingel, 2017

auseinandersetzen. In den letzten Jahrzehnten haben insbesondere die demokratischen Nationen viel daran gearbeitet, dass schädigende Vorurteile bestraft und ihre Auswirkungen Stück für Stück gemildert wurden. Auf der anderen Seite handelt es sich bei Erfahrungswissen immer ein Stück weit um Vorurteile im konkreten Sinne des Wortes, nämlich um die Erfahrung, dass manche Situationen eher zu diesem, andere wiederum zu einem anderen Ergebnis führen werden, um eine beobachtete Korrelation zwischen zwei Eigenschaften herzustellen, die keine direkte Kausalität beinhalten muss. Im besten Fall sind diese Vorurteile flexibel und geben nur ein erstes Indiz zur Beurteilung einer Situation, die sich danach einem strukturierten Entscheidungsfindungsprozess unterwirft. Bei algorithmischen Entscheidungssystemen haben wir es dagegen mit »algorithmisch legitimierten Vorurteilen« zu tun, die bei einem fertig konstruierten System erst einmal nicht flexibel sind: Die Entscheidungen sind durchdefiniert, von den einmal getroffenen Regeln kann es keine Abweichungen geben - die Vorurteile sind fest im Code eingeschrieben. Um diesem Problem zu begegnen, kann man zum Beispiel die Systeme ständig weiterlernen lassen. Dies geschieht beispielsweise bei allen Algorithmen der sozialen Medien, die unsere Nachrichten vorfiltern: Wir klicken jene Inhalte davon an, die wir tatsächlich sehen wollen und geben daher ein Feedback über die für uns getroffene Auswahl. Dadurch sind diese Systeme auch derart erfolgreich.

Dies ist allerdings nicht bei allen Systemen so einfach. So kann das System, das über die Kreditwürdigkeit eines Menschen entscheidet und ihn positiv evaluiert, zwar prinzipiell auch seine Gewichtungen verändern, wenn es darüber informiert wird, dass die Person den Kredit doch nicht, wie erwartet, zurückgezahlt hat. Aber an diesem Beispiel sieht man auch, dass das Feedback asymmetrisch sein kann: Die Person, die keinen Kredit bekommen hat, kann prinzipiell nicht

nachweisen, dass sie ihn zurückgezahlt hätte. Es ist bisher unklar, wie ein asymmetrisches Feedback auf lernende algorithmische Systeme wirkt Ein ständiges Weiterlernen hilft also nur bedingt dabei, dass die einmal getroffenen Vorurteile flexibel bleiben.

Ein weiterer wichtiger Aspekt ist, dass die Entscheidung, welche Menschen durch den Algorithmus als zueinander »ähnlich« angesehen werden, durch viele, einzelne Entscheidungen einer Vielzahl von Menschen beeinflusst wird. Dies sind unter anderem:

1. Die Datenauswahl, mit der die Menschen beschrieben werden;
2. Die Art und Weise, mit der schwer fassbare soziale Phänomene quantifiziert werden, z. B. die »Zuverlässigkeit« eines Mitarbeiters oder die Sozialprognose eines Straffälligen;
3. Die gewählte Methode des maschinellen Lernens;
4. Das gewählte Qualitäts- und Fairness-Maß;
5. Die Art und Weise, in der das algorithmische Entscheidungssystem eingesetzt wird, z. B. die Schulung der dateneingebenden Mitarbeiter oder das Training der letztendlich entscheidenden Person, wenn das Ergebnis des algorithmischen Entscheidungssystems nur entscheidungsunterstützend eingesetzt wird. Ein Beispiel dafür ist COMPAS - dem oben genannten Rückfälligkeitsvorhersagesystem -, das in vielen Gerichtsälen vor der Urteilsverkündung hinzugezogen wird.

Es ist also offensichtlich, dass diese Schritte jeweils qualitätsgesichert erfolgen müssen, um eine sinnvolle Interpretierbarkeit und Interpretation einer maschinellen Entscheidung überhaupt zu gewährleisten. Diesen Prozess aufzusetzen, bedarf des interdisziplinären Austausches und der gesellschaftlichen Diskussion, zum Beispiel über die Frage, ab welcher möglichen Schadenstiefe für Indivi-

duen oder Gesellschaft die Einhaltung der jeweiligen Qualitätskriterien notwendig ist und mit welchem Aufwand diese Einhaltung durchgesetzt werden soll.

3.2. Notwendige Diskurse im spezifischen Kontext

Bei der Frage nach den jeweils geeigneten Qualitäts- und Fairnessmaßen reicht eine allgemeine Diskussion nicht aus, um die Qualität eines algorithmischen Entscheidungssystems zu sichern, da diese Wahl stark kontextabhängig ist. Somit kann diese Wahl nicht ohne die Einbettung in ein Gesamtsystem, in dem die durch sie vorbereiteten oder getroffenen Entscheidungen ihren Sinn erhalten, evaluiert werden. Daher erscheint es hier insbesondere notwendig, dass - je nach möglicher Schadenstiefe für Individuum und Organisation - die betroffenen Personen oder die sie vertretenden Institutionen in den Festlegungsprozess mit eingebunden werden. Neben allgemeinen Diskriminierungsverboten sind hier - wie oben dargelegt - die Einbettung des algorithmischen Entscheidungssystems in den Gesamtkontext, aber auch die Kosten für falsch-positive und falsch-negative Entscheidungen. Diese Kosten müssen im jeweiligen Kontext definiert werden.

In vielen konkreten Einsatzszenarien werden dabei einander widersprechende Maße für die verschiedenen Aspekte einer fairen und hochwertigen Entscheidung unvermeidbar sein. Dies ist für sich genommen ein weiteres Indiz dafür, dass die jeweilige Situation sich grundsätzlich nicht für den Einsatz eines algorithmischen Entscheidungssystems eignet, da dieses übersimplifizierend agiert und damit der Komplexität der Lage nicht gerecht wird.

4. Fazit

In diesem Artikel haben wir gezeigt, dass es verschiedene Diskriminierungs- oder »Fairness«-Maße gibt, die jeweils ihre Berechtigung haben, sich aber gegenseitig ausschließen. Es ist interessant, dass dieser mathematische Widerspruch nicht früher entdeckt wurde, da auch menschliche Expertinnen und Experten sich entscheiden müssen, welche Art von Fairness und Diskriminierungsfreiheit sie verfolgen. Hier hat also der datenzentrierte Ansatz der Entwicklung eines algorithmischen Entscheidungssystems dazu geführt, dass ein grundsätzlicher Widerspruch in Entscheidungsprozessen aller Art aufgedeckt wurde. Die Gesellschaft und jeder einzelne Experte und jede einzelne Expertin muss sich dementsprechend entscheiden, welches Konzept ihm oder ihr wichtiger ist oder wie diese Aspekte zu gewichten sind. Ein algorithmisches Entscheidungssystem kann aber während des Trainings grundsätzlich immer nur bezüglich eines Qualitätsmaßes bewertet werden. Daher könnte das Fazit dieser Forschung sein, dass grundsätzlich mehrere, von unterschiedlichen Teams entwickelte Systeme zur Entscheidungsunterstützung verwendet werden sollten, um die Vielfalt an Meinungen dazu gebührend zu berücksichtigen. Dieser Befund sollte also Konsequenzen sowohl für die Ausbildung menschlicher Experten als auch für das Design und den Kauf von algorithmischen Entscheidungssystemen haben - insbesondere in der öffentlichten IT mit ihrem großen Einfluss auf das Gemeinwohl.

Quellen

ACLU (American Civil Liberty Union) (2011). *Smart Reform Is Possible - States Reducing Incarceration Rates and Costs While Protecting Communities*, Report from August 2011. http://s.fhg.de/BQN, abgerufen am 22.02.2018

Angwin, J., Larson, J., Mattu, S. & Kirchner, L. (2016). Machine Bias -There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*. 23.05.2016. http://s.fhg.de/ZS5, abgerufen am 25.01.2018

Ariely, D. (2010*). Predictably Irrational, Revised: The Hidden Forces That Shape Our Decisions*. HarperCollins, New York

Assheuer, T. (2017). Die Big-Data-Diktatur. *Die Zeit,* 30.11.2017

Danziger, S., Levav, J. & Avnaim-Pesso, L (2011). Extraneous factors in judicial decisions. *Proceedings of the National Academy of the Sciences* , 108 (S. 6889-6892)

Dieterich, W., Mendoza, C. & Brennan, T. (2016*). COMPAS risk scales: Demonstrating accuracy equity and predictive parity*. Northpointe Inc. http://s.fhg.de/B9V, abgerufen am 22.02.2018

EPIC (2017). *Algorithms in the criminal justice system*. http://s.fhg.de/7QT, abgerufen am 25.01.2018

Harcourt, B. E. (2006). *Against prediction: Profiling, policing, and punishing in an actuarial age*. University of Chicago Press

Helbing, D., Frey, B. S., Gigerenzer, G., Hafen, E., Hagner, M., Hofstetter, Y., van den Hoven, J., Zicari, R. V., Zwitter, A. & Kahneman, D. (2017*).* Das Digital-Manifest. In: Könneker, C. (Hrsg.), *Unsere digitale Zukunft: In welcher Welt wollen wir leben?*

Jaume-Palasí, L. & Spielkamp, M. (2017). Ethik und algorithmische Prozesse zur Entscheidungsfindung oder -vorbereitung. AlgorithmWatch Arbeitspapier Nr. 4. http://s.fhg.de/qPb, abgerufen am 22.02.2018

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016*).* Inherent trade-offs in the fair determination of risk scores. arXiv preprint arXiv:1609.05807

Krafft, T. D. & Zweig, K. A. (2018). Wie Gesellschaft algorithmischen Entscheidungen auf den Zahn fühlen kann. In: Resa Mohabbat Kar, Basanta E. P. Thapa & Peter Parycek (eds.). *(Un)berechenbar? Algorithmen und Automatisierung in Staat und Gesellschaft*. Kompetenzzentrum Öffentliche IT.

Lischka, K., & Klingel, A. (2017). *Wenn Maschinen Menschen bewerten*. Bertelsmann Stiftung.

Lischka, K. & Stöcker, C. (2017). *Digitale Öffentlichkeit – Wie algorithmische Prozesse den gesellschaftlichen Diskurs beeinflussen.* Bertelsmann Stiftung.

Massoumnia, M.-A., Verghese, G. C. & Willsky, A. S. (1989*).* Failure detection and identification. *IEEE transactions on automatic control,* 34(3), S. 316-321

Reindl, C. & Krügl, S. (2017). *People Analytics in der Praxis*. Haufe-Lexware

Robinson, D. & Yu, H. (2014). Knowing the Score: New Data, Underwriting, and Marketing in the Consumer Credit Marketplace. http://s.fhg.de/x8c, abgerufen am 22.02.2018

Schramm, U. & Schramm, H. (1990). *Automatische Sichtprüfung von Oberflächen mit neuronalen Netzen*. In: *Mustererkennung 1990:* 12. DAGM-Symposium Oberkochen-Aalen, 24.–26. September 1990, Vol. 254, S. 114. Springer.

Statista. (2018*). Länder mit der größten Anzahl an Inhaftierten pro 100.000 Einwohner (Januar 2018*)*. http://s.fhg.de/mk9, abgerufen am 25. 01.2018

Stiller, S. (2015). *Planet der Algorithmen*. Ein Reiseführer, München.

The Intercept (2015). *SKYNET: Courier Detection via Machine Learning*. http://s.fhg.de/me2, abgerufen am 22.2.2018

The Sentencing Project (2013). *Report of The Sentencing Project to the United Nations Human Rights Committee Regarding Racial Disparities in the United States Criminal Justice System.* http://s.fhg.de/izs, abgerufen am 25.02.2018

Zweig, K. A. (2016a). 1. Arbeitspapier: Was ist ein Algorithmus? http://s.fhg.de/sHL, abgerufen am 01.02.2018

Über die Autoren

Katharina A. Zweig

Prof. Dr. Katharina Zweig ist Professorin für theoretische Informatik an der TU Kaiserslautern und leitet dort das *»Algorithm Accountability Lab«.* Sie ist auch verantwortlich für den Studiengang Sozioinformatik an der TU Kaiserslautern. 2014 wurde sie zu einem von Deutschlands »Digitalen Köpfen« gewählt und 2017 bekam sie den ars-legendi-Preis in Informatik und Ingenieurswissenschaften des

4ING und des Stifterverbandes für das Design des Studiengangs Sozioinformatik.

Professorin Zweigs Forschungsinteresse liegt bei der Interaktion von IT-Systemen und Gesellschaft sowie der Analyse komplexer Netzwerke. Momentan bewertet sie, wie stark Algorithmen diskriminieren können und ob Google's Suchmaschinenalgorithmus Filterblasen erzeugt - dazu hat sie das Datenspendeprojekt federführend entwickelt und zusammen mit AlgorithmWatch und mit einer Förderung der Landesmedienanstalten durchgeführt. Sie berät zu diesen Themen Landesmedienanstalten, Gewerkschaften, Politik und Kirchen und ist Mitgründerin der Nichtregierungsorganisation AlgorithmWatch. Sie ist seit 2014 Mitglied im Innovations- und Technikanalyse-Beraterkreis des Bundesministeriums für Bildung und Forschung.

Tobias D. Krafft

Tobias D. Krafft ist Doktorand am Lehrstuhl *»Algorithm Accountability«* von Professorin Katharina A. Zweig an der TU Kaiserslautern. Als Preisträger des Studienpreises 2017 des Forums Informatiker für Frieden und gesellschaftliche Verantwortung reichen seine Forschungsinteressen von der (reinen) Analyse algorithmischer Entscheidungssysteme bis hin zum Diskurs um deren Einsatz im gesellschaftlichen Kontext. Im Rahmen seiner Promotion hat er das Datenspendeprojekt mit entwickelt und einen Teil der Datenanalyse durchgeführt. Er ist einer der Sprecher der Regionalgruppe Kaiserslautern der Gesellschaft für Informatik, die es sich zur Aufgabe gemacht hat, den interdisziplinären Studiengang der Sozioinformatik (TU Kaiserslautern) in die Gesellschaft zu tragen.