Wie Gesellschaft algorithmischen Entscheidungen auf den Zahn fühlen kann

Tobias D. Krafft & Katharina A. Zweig

Technische Universität Kaiserslautern

Zunehmend treffen algorithmische Entscheidungssysteme (ADM-Systeme) Entscheidungen über Menschen und beeinflussen damit öffentliche Räume oder die gesellschaftlichen Teilhabemöglichkeiten von Individuen; damit gehören derartige Systeme zur öffentlichen IT. Hier zeigen wir, am Beispiel der Analyse von Rückfälligkeitsvorhersagesystemen und dem Datenspende-Projekt zur Bundestagswahl 2017, wie solche Systeme mit Hilfe von Black-Box-Analysen von der Öffentlichkeit untersucht werden können und wo die Grenzen dieses Ansatzes liegen. Insbesondere bei ADM-Systemen der öffentlichen Hand zeigt sich hierbei, dass eine Black-Box-Analyse nicht ausreichend ist, sondern hier ein qualitätsgesicherter Prozess der Entwicklung und Evaluation solcher Systeme notwendig ist.

1. Einleitung

Die zunehmende Digitalisierung eröffnet der Gesellschaft die Möglichkeit, immer komplexere soziale Probleme durch *algorithmische Entscheidungssysteme,* sogenannte ADM-Systeme *(Algorithmic Decision Making Systems)*, anzugehen. Ein algorithmisches Entscheidungssystem ist dabei ganz allgemein ein durch einen Computer ausführbares Programm, das Menschen oder Objekten eine Bewertung zuweist, basierend auf einer Reihe von Eigenschaften des Subjektes oder des Objektes. Diese Bewertung kann ein »Risiko« für eine

bestimmte Handlungsweise oder eine Wahrscheinlichkeit für eine zukünftige Verhaltensweise darstellen.

Definition ADM-System

Algorithmische Entscheidungssysteme (Algorithmic Decision Making Systems - ADM-Systeme, die) enthalten eine algorithmische Komponente, die - basierend auf der Eingabe - eine Entscheidung bzgl. eines Sachverhaltes trifft, d. h., die einen einzigen Wert berechnet. Wenn der Algorithmus von Experten erarbeitet wurde, spricht man von einem Expertensystem. Daneben gibt es solche, die das Regelsystem mit Hilfe von maschinellem Lernen aus Daten selbstständig ableiten.

Aktuell werden bereits die Kreditwürdigkeit von Bankkunden,[1] die Unterstützungswürdigkeit von Arbeitslosen[2] oder die zukünftige Rückfallwahrscheinlichkeit von Straftätern in den USA algorithmisch bestimmt. Zu den ADM-Systemen gehören aber auch solche, die Nutzer eines internetbasierten Dienstes klassifizieren, um ihnen beispielsweise personalisiert weitere Produkte und Dienste anzubieten. So sind alle Arten von Produktempfehlungssystemen, aber auch der NewsFeed-Algorithmus von Facebook, die Auswahl der angezeigten Tweets auf Twitter oder die personalisierte Suchergebnisliste auf Googles Suchmaschine im Sinne der angeführten Definition. ADM-Systeme können auch Entscheidungen über Objekte treffen, z. B. im Bereich der Bilderkennung oder Qualitätsüberprüfung von Produkten. Im Rahmen dieses Artikels betrachten wir jedoch

[1] Lischka & Klingel, 2017

[2] Niklas, Sztandar-Sztanderska & Szymielewicz, 2015

nur solche ADM-Systeme, die Menschen anhand ihres jetzigen Handelns klassifizieren und/oder basierend auf einer solchen Klassifikation ihr zukünftiges Handeln vorhersagen, da diese Systeme für den Bereich der öffentlichen IT von besonderem Interesse sind.

Solche ADM-Systeme entspringen hochkomplexen soziotechnischen Designprozessen, welche von der Selektion der besten Datenquellen über die Auswahl der wissenschaftlichen Analysemethode bis hin zur bestmöglichen Visualisierung der Ergebnisse reichen. Da sich jedoch Anzeichen von Fehlern bei ADM-Systemen mehren[3], weil sie auf falschen Modellannahmen beruhen können und so beispielsweise ihre Ergebnisse sexistisch[4] oder rassistisch[5] sind, wächst der gesellschaftliche Wunsch nach Transparenz, geäußert unter anderem vom Justizminister des Saarlandes Stephan Toscani[6], von der rheinland-pfälzischen Ministerpräsidentin Malu Dreyer[7] und von Bundesjustizminister Heiko Maas[8].

Allerdings haben die Reaktionen auf die Veröffentlichung des PageRank-Algorithmus[9] von Google im Jahre 1998 ein zentrales Problem der Transparenzbemühungen offenbart. Diese Veröffentlichung führte umgehend dazu, dass Personen und Firmen erfolgreich Strategien entwickelten, um ihre Webseiten weit höher zu platzieren als dies im Sinne der Algorithmendesigner war. Insbesondere zählten dazu: Die Organisation von Linkfarmen, unsichtbare Platzierungen von beliebten Suchbegriffen oder der Ankauf von Links von Seiten

3 Zweig, Fischer & Lischka, 2018
4 Datta, Tschantz & Datta, 2015
5 Sweeney, 2013
6 Toscani, 2017
7 Dreyer, 2017
8 Maas, 2017
9 Brin & Page, 1998

mit hoher Glaubwürdigkeit (Stichwort: *Black Hat Search Engine Optimization*). Die hohe Transparenz über den Entscheidungsmechanismus des Algorithmus führte also zu einer den ursprünglichen Nutzen korrumpierenden Manipulation der Suchergebnisreihenfolgen. Damit ist offensichtlich, dass alle Designteams von algorithmischen Entscheidungssystemen, insbesondere von Recommendation-Systemen (»Empfehlungssystemen«) wie Suchmaschinenalgorithmen, Nachrichtenaggregatoren oder Produktempfehlungssystemen, eine Balance finden müssen zwischen öffentlichen Informationen über das System und der Manipulationssicherheit.

Eine Transparenz im Sinne der Publikation von Code ist aber auch in vielen Fällen gar nicht notwendig: So skizzierte Nicholas A. Diakopoulos Szenarien, die es zulassen, eine Black-Box-Analyse durchzuführen. Eine *Black-Box-Analyse* ist im Wesentlichen ein naturwissenschaftlicher Zugang zur Untersuchung von ADM-Systemen.[10] Hierbei wird das System als *Black Box*, also als geschlossene Schachtel betrachtet, in die man keinen direkten Einblick erhält. Dennoch kann man versuchen, einen Überblick zu bekommen, indem man diese mit Daten füttert und aus der Beziehung von Ein- und Ausgabe Rückschlüsse über die innenliegende Mechanik sowie über die Güte des Systems zieht.

In unserer Ausarbeitung beschreiben wir zwei Beispiele für solche Black-Box-Analysen für algorithmische Entscheidungssysteme aus dem Bereich der öffentlichen IT. Weshalb die zum Verständnis der Thematik benötigten technologischen Grundlagen in Abschnitt 2 ausgeführt werden. Beim ersten Beispiel haben wir die Qualität eines ADM-Systems anhand von öffentlich einsehbaren Beziehungen zwischen Eingabe, den durch die Maschine vorhergesagtem sowie wirklichem menschlichen Verhalten analysiert (Abschnitt 3). Dies

[10] Diakopoulos, 2014

ist möglich da das gesellschaftlich gewünschte Verhalten klar bekannt ist. In Abschnitt 4 zeigen wir dann eine Analysemöglichkeit für Ergebnisse von ADM-Systemen, bei denen nicht bekannt ist, was die korrekte oder beste Ausgabe gewesen wäre. Die Google-Suche stellt eine solche Situation dar, da hier die Grundwahrheit, also was genau die richtigen Ergebnisse in der richtigen Reihenfolge sind, weder im großen Ganzen noch speziell im individuellen Fall bekannt ist. In Abschnitt 5 werden abschließend aus den Ergebnissen relevante Forderungen abgeleitet.

2. Technologische Grundlagen

Algorithmische Entscheidungssysteme sind erst einmal solche Systeme, die anhand einer vorher festgelegten Regelbasis eine Reihe von Informationen über eine Situation oder Person verarbeiten und dann mit einem einzigen Berechnungsergebnis enden. Sehr einfache algorithmische Entscheidungssysteme sind z. B. Kreditwürdigkeitsverfahren bei Banken oder Einstufungsverfahren beim Abschluss einer Autoversicherung. In den letzten Jahren werden aber vermehrt algorithmische Entscheidungssysteme verwendet, welche die eigentlichen Entscheidungsregeln direkt aus Daten ableiten und zwar mit Hilfe von Methoden des maschinellen Lernens.[11] Die Vorgehensweise erklären wir am Beispiel von sogenannten »Rückfälligkeitsvorhersagealgorithmen«: Das US-amerikanische Justizsystem verwendet zum Beispiel schon seit Längerem in einigen Bundesstaaten

[11] Die Abgrenzung des maschinellen Lernens zur künstlichen Intelligenz sind nicht scharf. Manchmal werden die Begriffe synonym verwendet oder das eine als eine Unterkategorie des anderen angesehen. Wir verwenden in diesem Artikel den Begriff »maschinelles Lernen« um klarzumachen, dass es sich hier nicht um ein im landläufigen Sinne des Wortes intelligentes System handelt.

das *›Correctional Offender Management Profile for Alternative Sanctions‹,* kurz *COMPAS Assessment Tool*[12] genannt, das für bereits verurteilte Straftäter eine Klassifizierung nach prognostiziertem Rückfallrisiko vornimmt.[13] Dazu wird für jeden Straftäter ein »Score« berechnet, der es erlaubt, die Personen nach ihrem vermeintlichen Risiko zu sortieren. Um dieses System zu erstellen, bekamen die Entwickler des algorithmischen Entscheidungssystemes eine große Menge von Informationen über Kriminelle aus den letzten Jahren, zusammen mit der Information, ob diese Kriminellen wieder rückfällig wurden, z. B. innerhalb eines Zweijahreszeitraums. Methoden des maschinellen Lernens sind auf solchen Daten in der Lage, diejenigen Informationen zu identifizieren, die mit dem vorherzusagenden Verhalten stark korrelieren – in diesem Fall also der Rückfälligkeit. Klassische Beispiele für Informationen, die stark mit der Rückfälligkeit korreliert sind, sind z. B. das Geschlecht und das Alter: Männer werden sehr viel häufiger wieder rückfällig als Frauen und jüngere Personen öfter als ältere[14].

[12] Diese Web-Applikation wurde ursprünglich vom *Northpointe Institute for Public Management Inc*. als automatisierte Entscheidungsunterstützung zur Bewertung von Straffälligen entwickelt und vertrieben. 1998 als »Breitband«-Bewertungstool konzipiert, kann es anhand verschiedener Fragebögen in 22 verschiedenen Bedürfnis- und Risikobereichen Prognosen über Individuen erstellen. So soll es den Hilfebedarf der Bewerteten erkennen und quantifizieren und bietet unter anderem die Möglichkeit, aus dem sogenannten *CORE Risk Assessment*-Fragebogen, der aus 137 verschiedenen Fragen besteht, sowohl Prognosen über das *›General Recidivism Risk‹* (Generelles Rückfallrisiko) als auch das *›Violent Recidivism Risk‹* (Rückfallrisiko für Gewalttaten) zu treffen. Siehe Northpointe, 2017; Angwin, Larson, Mattu & Kirchner, 2016.

[13] Northpointe, 2017

[14] Florida Department of Corrections Recidivism Report, 2012

Die Methoden des maschinellen Lernens sind dabei in der Lage, auch kleine Unterschiede in der Rückfälligkeitsrate zwischen verschiedenen Subgruppen in den Daten zu berücksichtigen. Zudem sind sie in der Lage, nach viel mehr verschiedenen Korrelationen zu suchen: Anstatt nur jede Information wie Alter und Geschlecht auf ihren Einzeleffekt zu untersuchen, können Maschinen auch größere Teilmengen von Informationen auf Korrelationen mit der Rückfälligkeit untersuchen. Die gefundenen Korrelationen werden gewichtet und in einer Entscheidungsregelstruktur abgespeichert. Wir gehen im Folgenden davon aus, dass das algorithmische System für jede Person eine Zahl berechnet, mit der Vereinbarung, dass eine höhere Zahl mit einer höheren Wahrscheinlichkeit für »Rückfälligkeit« einhergeht. Die *American Civil Liberty Union* schlägt vor, auf einem solchen Scoring basierend alle Verdächtigen in drei Risikoklassen einzuteilen (s. Abbildung 2): Personen mit hohem, mittlerem und niedrigem Risiko.[15] Dafür werden zwei Schwellwerte bestimmt, sodass Personen mit einem Score über dem höchsten Schwellwert in der Hochrisikoklasse sind, solche mit einem Wert unter dem niedrigeren Schwellwert in der Niedrigrisikoklasse und alle anderen in der Klasse mit mittlerem Risiko. In diesen Klassen sind nun Personen, die entweder wieder rückfällig wurden oder nicht. Der Anteil der rückfälligen Personen in jeder Klasse wird zum Schluss dann als Rückfälligkeitsrisiko jedes Individuums in der Klasse interpretiert.

In Abbildung 2 sind in der Hochrisikoklasse insgesamt 10 Personen, von denen 5 rückfällig wurden: Jede weitere Person, die hier eingeordnet wird, bekommt also ein individuelles Risiko von 50 Prozent zugewiesen. Personen, die vom Algorithmus in die Klasse mit »mittlerem Risiko« eingeordnet werden, bekommen dementsprechend ein individuelles Risiko von 30 Prozent zugewiesen und solche in der

[15] ACLU, 2011

verbleibenden Klasse eins von nur 20 Prozent. Es ist offensichtlich, dass diese Kategorisierung nicht optimal ist: Im Bestfalle wären alle Rückfälligen in der Hochrisikoklasse und alle anderen in der Niedrigrisikoklasse, die dann eine »Nullrisikoklasse« wäre.

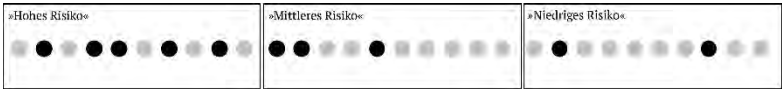

Abbildung 2: Algorithmische Entscheidungssysteme sortieren und klassifizieren Datenpunkte, hier dargestellt als Kreise. Schwarze stellen hier beispielsweise Personen dar, die bekanntermaßen rückfällig wurden, und graue solche, die das nicht wurden. Ein Scoring und eine darauf basierende Sortierung wäre optimal, wenn alle »Rückfälligen« (schwarze Kreise) auf der einen Seite stünden (z. B. links) und alle »Nicht-Rückfälligen« auf der anderen – dann könnte man die beiden Personengruppen leicht voneinander unterscheiden. Eine solche Sortierung – basierend auf den bekannten Eigenschaften der Personen – ist allerdings oftmals nicht möglich. Die durch den Algorithmus berechnete Sortierung der bekannten Daten dient im Beispiel der Rückfälligkeitsvorhersage als Grundlage für eine Klassifikation in »Risikoklassen«.

Die Qualität der gefundenen Entscheidungsregeln wird nun getestet, indem die Maschine einen neuen Datensatz von Kriminellen vorgelegt bekommt, bei dem zwar dem Tester bekannt ist, welche Person wieder rückfällig wurde, aber nicht der Maschine. Dieses bewertet den neuen Datensatz anhand der gefundenen Entscheidungsregeln. Das Ergebnis kann auf verschiedene Arten und Weisen in seiner Qualität bewertet werden. Wir gehen im Folgenden auf die Sinnhaftigkeit der Verwendung zweier möglicher Qualitätsmaße ein, nämlich die ROC AUC und die PPVk, als Beispiel für eine Fragestellung, die Gesellschaft mit Hilfe einer Blackbox-Analyse beantworten kann.

3. Analyse der Qualität von algorithmischen Entscheidungssystemen

Wird ein ADM-System eingesetzt, das zur öffentlichen IT zählt, sollte sichergestellt sein, dass seine Auswirkungen tatsächlich im Sinne des Gemeinwohls sind, dass sie qualitativ hochwertig sind und dass sie keine gesellschaftlichen Werte und Übereinkommen verletzen. Bei diesen - meist proprietären - Systemen muss die Öffentlichkeit jedoch aufgrund der geringen Transparenz der Algorithmen und der daraus resultierenden fehlenden Einsicht in den Bewertungsprozess häufig den Evaluierungsprozessen der Firmen oder staatlicher Institutionen blind vertrauen. Das oben schon genannte Rückfälligkeitsvorhersagesystem COMPAS beispielsweise gibt an, dass seine Qualität bei 70 Prozent läge. Wir zeigen mit diesem Beispiel, wie mit Hilfe von veröffentlichten Eingabedaten und der entsprechenden Ausgabe eines ADM-Systems und dem Wissen um das wirkliche Verhalten der Beurteilten nachgewiesen werden konnte, dass dieses ADM-System sowohl zu wenig hochwertig ist, als auch diskriminierende Entscheidungen trifft.

Der von ProPublica[16] veröffentlichte Datensatz[17] enthielt über eine Reihe von Kriminellen wesentliche Informationen, den von COMPAS berechneten Scoring-Wert, die darauf beruhende Klassifikation der Personen und die Antwort auf die Frage, ob diese Person wieder rückfällig geworden war. Eine Analyse dieses Datensatzes zeigte zuerst einmal, dass der von der Firma genannte Prozentsatz an »korrekten Entscheidungen« auch für diesen Datensatz gilt. Das heißt, die genannten 70 Prozent an korrekten Entscheidungen konnten

[16] ProPublica ist eine durch Spenden finanzierte, US-amerikanische Journalisten- und Rechercheplattform.

[17] http://github.com/propublica/compas-analysis

auch auf dieser Menge von Kriminellen reproduziert werden. Im Folgenden skizzieren wir eines unserer Forschungsergebnisse, das aufweist, dass das von der Fima verwendete Qualitätsmaß schnell sehr hohe Werte erreicht, die aber für den realen Prozess, in dem die Systeme eingesetzt werden, wenig aussagekräftig sind.[18] Das von der Firma verwendete Qualitätsmaß ist die sogenannte *Receiver-Operator Characteristic Area under the curve* (ROC AUC). Diese ist eines der populärsten Qualitätsmaße im Bereich des maschinellen Lernens und gibt an, wieviele Paare von »Rückfälligen« und »Nicht-Rückfälligen« durch das System korrekt sortiert werden:[19] Ein Paar gilt als korrekt sortiert, wenn die rückfällig gewordene Person vom System einen höheren Wert zugemessen bekommt als die Person, die nicht rückfällig geworden ist. In Abbildung 2 liegt der Anteil der korrekt sortierten Paare (schwarzer Punkt liegt links von grauem Punkt) bei genau 70 Prozent ROC AUC. d. h., von den insgesamt 200 schwarzgrauen Paaren sind 140 korrekt sortiert. Demgegenüber stand bei unserer Forschung der *Positive Predictive Value* (PPVk), das ist der Anteil an rückfälligen Personen unter den Personen mit den k höchsten Scoring-Werten, wobei k der (bekannten) Anzahl von rückfälligen Personen im Datenset entspricht. Im obigen Beispiel sind bekanntermaßen 10 Rückfällige (10 schwarze Kugeln), aber unter den 10 Kugeln mit den höchsten Werten (den 10 am weitesten links stehenden Kugeln) befinden sich nur 5 Rückfällige. Damit liegt der PPVk-Wert bei 50 Prozent.

Die ROC AUC ist ein sinnvolles Qualitätsmaß, wenn es nur zwei Personen zur Bewertung gibt und eine davon gewählt werden muss – wenn zum Beispiel zwei Kandidaten sich auf eine Stelle bewerben, die sofort besetzt werden muss. Vor Gericht stehen wir aber in einer

[18] Krafft, 2017a

[19] Hanley & McNeil, 1982

anderen sozialen Situation: Bei jeder Angeklagten und jedem Angeklagten müssen die Richter entscheiden, ob es genügend Anzeichen für ein hohes Rückfälligkeitsrisiko gibt – es handelt sich also eher um eine Schwellwertbetrachtung, wie sie auch dem PPVk zu Grunde liegt. Dabei können Richterinnen und Richter durchaus beurteilen, wie hoch das »k« normalerweise ist: Sie haben jahrelange Statistiken darüber, wieviel Prozent der Kriminellen wieder rückfällig werden. Damit ist nachvollziehbar, dass die ROC AUC und die PPVk zwei unterschiedliche Situationen bewerten: Nämlich, ob ein Algorithmus in der Lage ist, für viele Paare von Kandidaten den jeweils rückfälligeren korrekt zu bewerten oder ob ein Algorithmus sehr viele mit hohen Risikowerten belegt, die tatsächlich rückfällig werden.

Es zeigt sich ganz allgemein, dass die ROC AUC für alle sozialen Prozesse, bei denen aus einer kleinen Auswahl von Personen die optimale Wahl getroffen werden soll, ein gut interpretierbares Maß ist. Geht es darum, aus einer Gesamtbevölkerung die Personen mit dem höchsten Risiko zu benennen, ist der PPVk-Wert aussagekräftiger.

Es wäre unproblematisch, wenn die beiden Werte immer ähnlich hoch wären, die ROC AUC also auch ungefähr angibt, wie hoch der PPVk ist. Diese direkte Beziehung zwischen dem ROC AUC und dem PPVk-Wert ist aber weder *notwendigerweise*[20] noch *erwartbarerweis***e** der Fall (aktuelle eigene Forschung). Wir konnten konkret an dem von ProPublica zu diesem Zweck veröffentlichten Datensatz zeigen, dass die Hochrisikoklasse für allgemeine Rückfälligkeit nur einen Anteil von um die 50 Prozent Rückfälligen enthält - ein Prozentwert, der deutlich unter den 70 Prozent der paarweise korrekt Sortierten liegt. Für schwere Straftaten (z. B. Körperverletzung, Raub) liegt die paarweise Korrektheit ebenfalls wieder bei 70 Prozent ROC AUC,

[20] Krafft, 2017a

aber in der Hochrisikoklasse sind nur 20 Prozent der dort einsortierten Personen wieder rückfällig geworden. Dieser geringe Prozentsatz an Rückfälligen verbietet es eigentlich, die entsprechende Gruppe als »Hochrisiko«-Klasse zu bezeichnen. Der starke Unterschied zwischen der hohen ROC AUC und dem kleinen PPVk wird durch die insgesamt deutlich kleinere Gruppe an Personen bedingt, die mit schweren Straftaten rückfällig werden. In Abbildung 3 werden 100 Datenpunkte gezeigt, von denen nur 10 Datenpunkte Personen darstellen, die rückfällig wurden (in schwarz), alle anderen wurden dies nicht. Das gezeigte Scoring mit den rückfälligen Personen auf Position 4, 5, 19, 20, 21, 35, 36, 59, 60, 66 hat ebenfalls eine paarweise gemessene Qualität von 70 Prozent ROC AUC. Die Größe der Hochrisikoklasse wird aufgrund der bisherigen Fallraten abgeschätzt und entspricht damit der absoluten Anzahl an rückfälligen Personen im Datensatz (im Beispiel also 10 Personen). Von den 10 Personen mit dem höchsten Score sind aber nur 20 Prozent rückfällig geworden.

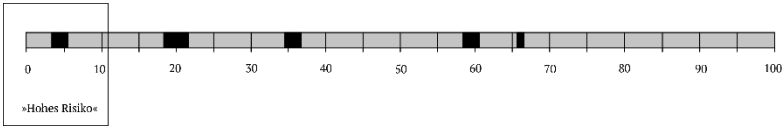

Abbildung 3: Bei höheren Ungleichgewichten zwischen Personen der beiden Klassen (z.B. rückfällig vs. nicht-rückfällig) kann ein hoher, paarweise korrekt sortierter Anteil leicht erreicht werden.

Damit zeigt sich eindeutig, dass das vom Unternehmen gewählte Qualitätsmaß für den sozialen Prozess, in den das ADM-System Anwendung findet, irreführend ist. Dies konnte ohne Offenlegung des Codes durch bloße Offenlegung der dafür notwendigen Testdaten analysiert werden. Insgesamt lässt sich schlussfolgern, dass ein von

öffentlicher Hand bezahltes und genutztes ADM-System einem hohen Qualitätsanspruch genügen muss, der nicht erst durch die Öffentlichkeit als unzureichend evaluiert werden sollte.[21]

Einen anderen Aspekt von ADM-Systemen konnte ProPublica mit diesem Datensatz ebenfalls untersuchen, nämlich die Frage nach einer eventuellen Diskriminierung. Das Team von Journalisten wies tatsächlich eine Ungleichbehandlung von Afroamerikanern und Weißen nach, indem auf Grundlage von 7000 in den Jahren 2013 und 2014 verhafteten Straftätern die jeweiligen Prognosen ausgewertet wurden.[22] Der daraufhin entbrannten Diskussion haben wir uns in einem weiteren Beitrag in diesem Band mit dem Titel »Fairness und Qualität algorithmischer Entscheidungen« gewidmet.[23]

Zusammenfassend lässt sich feststellen, dass neben der reinen Bewertung, wie »gut« eine Klassifikation durch ein solches ADM-System ist, auch andere gesellschaftliche Aspekte wie Diskrimierung und Fairness quantifiziert und für die Öffentlichkeit nachvollziehbar und unter Umständen nachprüfbar dokumentiert werden müssen. Dies kann gelingen, wenn es um ein vorherzusagendes Verhalten geht, dass direkt messbar ist, wenn der Entscheidung des Algorithmus also die »Wahrheit« gegenüber gestellt werden kann.

4. Analyse des Personalisierungsgrades von personalisierten, algorithmischen Entscheidungssystemen

Als deutlich schwieriger stellen sich die Fälle heraus, in denen unklar ist, was die bestmögliche Ausgabe der Maschine gewesen wäre.

[21] Zweig, Wenzelburger & Krafft, 2018

[22] Angwin et al., 2016

[23] Zweig & Krafft, 2018

Dies geschieht insbesondere bei allen personalisierten Onlinediensten: Dies sind ADM-Systeme, die – basierend auf Big Data – zuerst eine feingranulare Klassifikation der Nutzer durchführen, die opak bleibt. Für den eigentlichen Dienst gibt der Nutzer oder die Nutzerin meist weitere Daten ein, beispielsweise eine Suchanfrage bei einer Suchmaschine, die, zusammen mit den opak errechneten »Persönlichkeitswerten« und der ebenfalls nicht bekannten Menge an möglichen Suchergebnissen eine personalisierte Suchergebnisliste ergibt. Eine explizite Eingabe von Informationen durch den Nutzer ist aber nicht immer notwendig: Der Facebook-Newsfeed ist ein solches Beispiel. Auch bei diesem, die Öffentlichkeit gestaltenden IT-Artefakt ist unklar, welche Daten für die Sortierung herangezogen werden und nach welchen Kriterien die Selektion erfolgt. Obwohl Facebook die Ziele des Newsfeeds bekannt gegeben hat,[24] ist dazu bis heute kein weiteres Statement veröffentlicht worden, welche Faktoren genau berücksichtigt werden, um die für den Nutzer vermeintlich relevanten Inhalte nach oben zu sortieren. In einem aktuellen Beitrag beschreibt Mark Zuckerberg, Gründer und Vorstandsvorsitzender von Facebook Inc., welchen Einfluss er auf die Sortierung nehmen möchte, um das neu gesteckte Ziel – bedeutungsvollere soziale Interaktionen[25] – zu verfolgen.[26] Wie das erreicht werden soll und welche Parameter dafür eine Rolle spielen, bleibt allerdings wieder vage. Die jeweilig verborgenen Eingaben und die unklare Definition des Desiderats führt dazu, dass wir als Gesellschaft nicht wissen, was eigentlich die »richtige« Ausgabe des ADM-Systems

[24] *»most relevant stories at the top«* http://s.fhg.de/37n

[25] *»Based on this, we're making a major change to how we build Facebook. I'm changing the goal I give our product teams from focusing on helping you find relevant content to helping you have more meaningful social interactions.«* Siehe Zuckerberg, 2018.

[26] Zuckerberg, 2018

wäre. Dies gilt schon, wenn wir für die Gesellschaft als Ganzes entscheiden müssten, was die relevantesten Nachrichten sind - damit haben die Redaktionskonferenzen in Deutschland jeden Tag zu kämpfen. Es wird deutlich erschwert, diese Frage zu beantworten, wenn unbekannte Eingabedaten dazu genutzt werden, die Ausgabe zu personalisieren. Eine mögliche Analyse besteht aber in der Beantwortung der Frage, wie groß das Ausmaß der Personalisierung ist.

4.1. Datenspende BTW17: Wie stark personalisiert Google?

Die Google Suche stellt als soziales Medium Öffentlichkeit her und ist somit als ein Teil der öffentlichen IT zu verstehen.[27] Um das Ausmaß der Personalisierung bei dieser bekannten Black Box zu analysieren, haben wir anlässlich der Bundestagswahl 2017 ein Datenspende-Projekt[28] mit einem Citizen-Science-Ansatz entwickelt. Gefördert von den Landesmedienanstalten Bayern, Berlin-Brandenburg, Hessen, Rheinland-Pfalz, Saarland und Sachsen sowie mit Spiegel Online als Medienpartner haben uns über 4000 Menschen zu festen Suchzeitpunkten die von Google und Google News ausgelieferten Ergebnisse zu 16 Suchbegriffen[29] eingeschickt, darunter die Suchbegriffe »Angela Merkel« und »Martin Schulz« und die wichtigsten Parteinamen. Somit stehen uns fast 6 Millionen Suchergebnisse zur Auswertung zur Verfügung. Wir fanden heraus, dass im Durchschnitt, wenn wir die Suchergebnisse zweier beliebiger Datenspender vergleichen, bei der Suche nach Politikernamen nur ein bis zwei

[27] Hoeppner et. al., 2016, S. 35

[28] https://datenspende.algorithmwatch.org

[29] Angela Merkel, Martin Schulz, Christian Lindner, Katrin Göring-Eckardt, Cem Özdemir, Sahra Wagenknecht, Dietmar Bartsch, Alice Weidel, Alexander Gauland, CDU, CSU, SPD, FDP, Bündnis90/Die Grünen, Die Linke, AfD

Links nicht auf beiden Listen stehen. Es handelt sich also um eine vergleichsweise geringe Personalisierung. Für Parteien betrug die Anzahl der uneinheitlichen Links im Durchschnitt drei bis vier, wovon allerdings ein bis drei Links regionaler Natur waren. D. h., ein Nutzer in Berlin bekam die Berliner SPD-Seite angezeigt, während die Münchnerin die SPD-Seite des Ortsvereins in München bekam.[30] Damit wird der von Eli Pariser 2011 postulierten Filterblasentheorie für die Suchmaschine von Google die Grundlage entzogen. Dieses Projekt ist als *Proof of Concept* zu betrachten, da wir erstmalig zeigen konnten, wie ein so wichtiger Aspekt einer öffentlichen IT von der Gesellschaft untersucht werden kann - gleichzeitig handelt es sich aber durch die freiwillige Teilnahme nicht um eine repräsentative Nutzerstichprobe. Zudem ist die Aussage beschränkt auf die untersuchten Suchbegriffe und den Zeitpunkt der Untersuchung. Eine Verstetigung dieser Kontrolle mit wechselnden Suchbegriffen und einer repräsentativen Nutzerstichprobe wäre sowohl wünschenswert als auch mit geringen Kosten durchführbar.

Es wurde damit eine Analysemöglichkeit geschaffen, mit der erstmalig ein solcher Algorithmus durch die Einbeziehung und Mithilfe der Bevölkerung auf relevante Phänomene untersucht werden kann, ohne den dahinterliegenden Code oder die korrekte Sortierung zu benötigen.

5. Fazit: Entwicklung und Evaluation von ADM-Systemen in der öffentlichen IT

Zunächst ist festzuhalten, dass algorithmische Entscheidungssysteme als Teil der öffentlichen IT in der heutigen technisierten und komplexen Welt in vielen Bereichen des gesellschaftlichen Lebens

[30] Krafft, Gamer, Laessing & Zweig, 2017b; Krafft, Gamer & Zweig, 2017c

bereits so fest verankert und integriert sind,[31] dass es höchste Zeit für eine auf breiter Ebene geführte Debatte ist, wie die Gesellschaft mit der Nutzung, der Anschaffung und der Evaluierung von ADM-Systemen umgehen will und kann.[32] Solche Systeme können ebenso große Chancen bieten, die Zukunft positiv zu gestalten, wie sie auch Gefahren für die Gesellschaft bergen und sollten daher nicht als unkontrollierte und undurchschaubare Black Boxes Verwendung finden.

Allerdings scheint es so, als sei die Brisanz dieser Frage noch gar nicht wirklich in der Gesellschaft angekommen, denn es fehlt der Allgemeinheit offensichtlich noch an den Grundvoraussetzungen, um in eine derartige Debatte einsteigen zu können. Wie 2015 von amerikanischen Wissenschaftlern herausgefunden wurde, waren sich in einer der wenigen Studien dazu über 60 Prozent der befragten Nutzerinnen und Nutzer von Facebook nicht darüber im Klaren, dass hinter dem Newsfeed eine algorithmische Kuratierung steckt.[33] Wenn nun aber über die Hälfte der Nutzer nicht einmal von der Existenz solcher ADM-Systeme weiß und wie diese mittlerweile wesentliche Bereiche ihres Lebens mitbestimmen, lässt sich auch keine problembewusste Debatte über einen adäquaten Umgang mit solchen Systemen führen. Daher finden auch Studien wie die des *Electronic Privacy Information Centre* (EPIC) wenig Widerhall, die auf eklatante Unterschiede im Evaluierungsprozess der in der Recht-

[31] Lischka & Klingel, 2017

[32] Zweig, Wenzelburger & Krafft, 2018

[33] Eslami et al., 2015. Es ist zu betonen, dass die Stichprobe mit 40 Personen sehr klein war und somit das Ergebnis mit einer hohen möglichen Varianz behaftet ist.

sprechung der US-amerikanischen Staaten genutzten algorithmischen Entscheidungssysteme hinwiesen.[34] Angesichts der unaufhaltsamen Weiterentwicklung der Technik besteht ein starkes Ungleichgewicht zum entsprechenden Kenntnis- und Informationsstand bei der breiten Bevölkerung, das zeitnah aufgearbeitet werden müsste. Wir haben hier aber auch aufgezeigt, dass es nicht notwendigerweise einer Offenlegung oder sonstigen Öffnung des Codes bedarf, um ADM-Systeme zu analysieren und somit zu wichtigen, gesellschaftlich relevanten Erkenntnissen bezüglich möglicher Folgen beim Einsatz vom ADM-Systemen zu kommen.

Daraus lassen sich grundsätzlich folgende Forderungen ableiten:

1. Aufgrund des Tempos der technischen Entwicklung bedarf es dringend weiterer Forschung zum Einsatz und zur Kontrolle von ADM-Systemen im Bereich der öffentlichen IT.
2. Es bedarf zeitnah einer effizienten Aufklärung der Öffentlichkeit über den Einsatz von ADM-Systemen, die im Bereich der öffentlichen IT über Individuen entscheiden, einschließlich der Chancen und Gefahren solcher Systeme und einer Sensibilisierung für mögliche Probleme. Hierzu sind eine intensive Zusammenarbeit und ein enger Austausch zwischen Wissenschaft, Politik, Medien und Schulen erforderlich.
3. Wir brauchen einen standardisierten, qualitätssichernden Prozess zur Entwicklung und dauerhaften Evaluation von ADM-Systemen in der öffentlichen IT.

[34] EPIC, 2017

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. (2916). 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

Brin, S. & Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems,* Bd. 30, Nr. 1–7, S. 107–117

Datta, A., Tschantz, M. C. & Datta, A. (2015). Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies,* 2015(1), S. 92-112

Diakopoulos N. (2014). *Algorithmic Accountability Reporting: On the Investigation of Black Boxes*, Tow Center, Feb.

Dreyer M. (2018). *Ministerpräsidentin Malu Dreyer: Mehr Transparenz bei automatisierten Entscheidungen durch Algorithmen.* http://s.fhg.de/2Yq, abgerufen am 22.02.2018

EPIC (2017). *Algorithms in the criminal justice system*. http://s.fhg.de/7QT, abgerufen am 25. 01.2018

Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Vuong, A., Karahalios, K., Hamilton, K. & Sandvig, C. (2015). I always assumed that I wasn't really that close to [her]: Reasoning about Invisible Algorithms in News Feeds. *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, S. 153-162.

Florida Department of Corrections Recidivism Report (2012). Florida Prison Recidivism Study Releases From 2003 to 2010. Florida Department of Corrections. http://s.fhg.de/E4R, abgerufen am 20.04.2018

Hanley, J. A. & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology.* 143, Nr. 1, S. 29–36

Hoepner, P., Weber, M., Tiemann, J., Welzel, C., Goldacker, G., Stemmer, M.,,Weigand, F.,Fromm, J., Opiela, N. & Henckel, L. (2016). *Digitalisierung des Öffentlichen*. Kompetenzzentrum Öffentliche IT, Berlin,

Krafft, T. D. (2017a). *Qualitätsmaße binärer Klassifikatoren im Bereich kriminalprognostischer Instrumente der vierten Generation*. Masterarbeit. TU Kaiserslautern. arXiv preprint arXiv:1804.01557.

Krafft, T. D. & Gamer, M. & Laessing, M. & Zweig, K.A. (2017b). Filterblase geplatzt? Kaum Raum für Personalisierung bei Google-Suchen zur Bundestagswahl 2017. AlgorithmWatch, Sep. 2017.

Krafft, T. D. & Gamer, M. & & Zweig, K. A. (2017c). Personalisierung auf Googles Nachrichtenportal während der Bundestagswahl 2017. AlgorithmWatch, Feb. 2018.

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

Maas, H. (2018). Zusammenleben in der digitalen Gesellschaft – Teilhabe ermöglichen, Sicherheit gewährleisten, Freiheit bewahren. Pressemitteilung. http://s.fhg.de/Ciu, abgerufen am 22. Februar 2018

Niklas, P. & Sztandar-Sztanderska, K. & Szymielewicz, K. (2015). Profiling the Unemployed in Poland: Social and Political Implications of Algorithmic Decision Making. Panoptykon Foundation, Warsaw, Poland

Northpointe Inc. (2017). COMPAS Risk & Need Assessment System - Selected Questions Posed by Inquiring Agencies. http://s.fhg.de/7FX, abgerufen 03/2017[35]

Pariser, E. (2011). *The Filter Bubble: What the Internet Is Hiding from You*. Penguin Press, New York.

Sweeney, L. (2013). Discrimination in online ad delivery. *Queue*, 11(3), 10.

Toscani, S. (2018). Justizminister Stephan Toscani: »Der Staat muss eine digitale Infrastruktur schaffen, die auch im Internet einen pluralen öffentli-

[35] Aufgrund der Namensänderung der Firma Northepointe Inc in *equivant* ist im Zuge des Umzugs des Webauftritts diese Ressource nicht korrekt verlinkt.

chen Raum und Grundrechtsschutz gewährleistet – Einsatz von Algorithmen und Social Bots regulieren.«. http://s.fhg.de/s36, abgerufen am 22.02.2018

Zuckerberg, M. (2018). One of our big focus areas for 2018. Facebook. http://s.fhg.de/V4z, abgerufen am 25.01.2018

Zweig, K.A., Fischer, S. & Lischka, K. (2018). Wo Maschinen irren können – Fehlerquellen und Verantwortlichkeiten in Prozessen algorithmischer Entscheidungsfindung. Bertelsmann Stiftung.

Zweig, K. A., Wenzelburger, G. & Krafft, T. D. (2018). On chances and risks of security related algorithmic decision making systems. *European Journal for Security Research*.

Über die Autoren

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.

Katharina A. Zweig

Professorin 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 Googles 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.