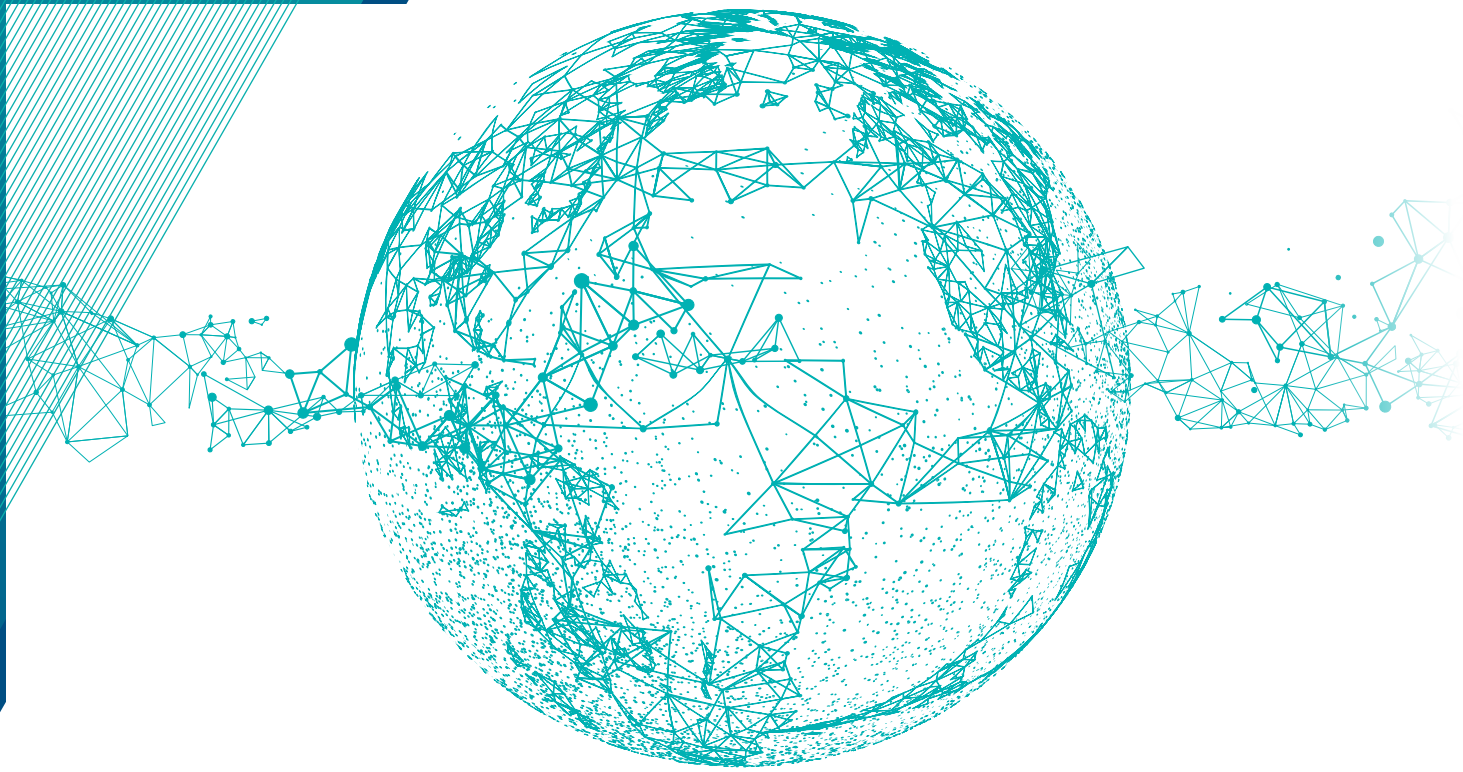
Nr. 338 / Januar 2019
Analysen & Argumente
Digitale Gesellschaft

KONRAD ADENAUER STIFTUNG



Algorithmische Entscheidungen: Transparenz und Kontrolle

Katharina A. Zweig

- Algorithmen sind Regeln zur Lösung mathematisch beschreibbarer Probleme. Sie sind ein wesentlicher Teil „Algorithmischer Entscheidungssysteme“ (ADM-Systeme), die in der Versicherungsbranche Risiken bewerten oder in anderen Kontexten vergleichbare Aufgaben lösen können.
- Wenn die Entscheidungsregeln für die Risikobewertung unklar sind, können selbstlernende ADM-Systeme verwendet werden. Diese lernen aus bisherigen Fällen und Daten.
- Allerdings lässt sich das Zustandekommen der Ergebnisse in der Regel nicht nachvollziehen. Selbstlernende ADM-Systeme bedürfen insbesondere immer dann einer speziellen Kontrolle, wenn sie Entscheidungen über Menschen treffen und Fehlentscheidungen einzelne Individuen bzw. die Gesellschaft als Ganzes schädigen können.
- Dazu reicht es nicht aus, Algorithmen alleine überprüfbar zu machen, vielmehr müssen die ADM-Systeme insgesamt (Algorithmen und Daten) und das sozio-informatische Gesamtsystem, in dem sie genutzt werden, in den Blick genommen werden.

Inhaltsverzeichnis

In der öffentlichen Debatte über Digitalisierung und KI ist der Begriff Algorithmus allgegenwärtig. Auf der einen Seite verkauft man Algorithmen als geradezu magische Lösung für schwierige Probleme. So bewirbt eine Firma ihre Software, die Angestellte in ihrer Leistung bewerten soll, mit den Worten: „In the end, with the availability of good data, the predictive possibilities are virtually unlimited." Auf der anderen Seite sind Algorithmen so sehr in Verruf geraten, dass neue Kontrollorgane und größere Einblicke in ihre Wirkungsweise gefordert werden (Meyer-Schöneberger & Cukier, 2013). Doch geraten die Begrifflichkeiten zuweilen durcheinander, wenn beispielsweise ein „Algorithmen-TÜV" verlangt wird, um Algorithmen zu kontrollieren, oder wenn Transparenz eines Algorithmus mit der Veröffentlichung seines Codes gleichgesetzt wird. Bei näherer Betrachtung zeigt sich, dass die Algorithmen, deren Überwachung gefordert wird, eigentlich nur einen kleinen Teil sogenannter *algorithmischer Entscheidungssysteme* ausmachen. Für diese zeigt die aktuelle Forschung allerdings, dass es nicht ausreicht, sie als einzelnes Produkt zu kontrollieren, sondern dass auch ihre Einbettung in unsere Gesellschaft der Überprüfung bedarf. Um diese Schlussfolgerung nachvollziehen zu können, ist es wichtig, die Begriffe „algorithmisches Entscheidungssystem" und „Algorithmus" zu differenzieren.

1. Algorithmische Entscheidungssysteme

Algorithmische Entscheidungssysteme (algorithmic decision making systems oder auch ADM-Systeme) beinhalten Regeln, nach denen eine Entscheidung getroffen werden kann. Dazu gehören einfache Systeme, wenn etwa KFZ-Versicherungen auf Grundlage von Daten Schadensfreiheitsklassen bestimmen. Die Entscheidungsregeln sind hier klar und für Menschen einsichtig: Sie beruhen im Wesentlichen auf dem Alter der Fahrer und ihrer bisherigen Unfallhistorie. In der aktuellen Diskussion geht es allerdings um Entscheidungssysteme, deren Entscheidungsregeln von Algorithmen selbständig abgeleitet werden.

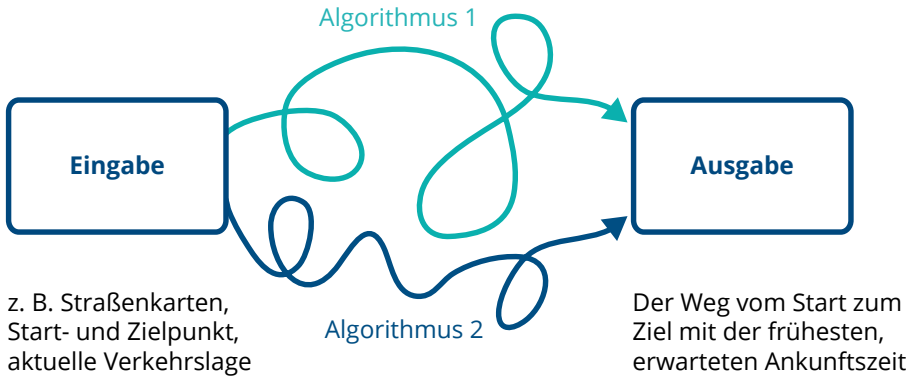
Etymologie des Begriffes „Algorithmus"
Der Begriff Algorithmus leitet sich von einem arabischen Mathematiker namens Al-Chwarizmi ab, der im 9. Jahrhundert ein mathematisches Lehrbuch verfasste. Es besteht Verwechslungsgefahr mit dem Logarithmus und mit Rhythmus, die aber etymologisch nicht mit dem Begriff „Algorithmus" verwandt sind.

Was sind Algorithmen?

Regeln zum Lösen von Problemen

Algorithmen finden eine Lösung für mathematisch beschreibbare Probleme, die in unterschiedlichen Anwendungssituationen immer wieder gelöst werden müssen. Ein mathematisches Problem benennt die vorliegenden Informationen (Eingabe) und definiert Eigenschaften, die eine auf den Informationen basierende Lösung haben soll (Ausgabe). Ein Beispiel ist die Berechnung des kürzesten Weges, basierend auf Kartenmaterial, dem Start- und Zielpunkt, u. U. um aktuelle Straßenverhältnisse ergänzt. Algorithmen sind Handlungsanweisungen, die zu einem Ergebnis mit den gewünschten Eigenschaften führen, basierend auf den eingegebenen Informationen. In der Regel gibt es mehrere Algorithmen, um zu einer Lösung zu gelangen (Abbildung 1).

Mathematisches Problem
Definiert das Verhältnis von Eingabe und Ausgabe



Algorithmus
Definiert eine Folge von Handlungen, die – basierend auf der Eingabe – eine Ausgabe mit den gewünschten Eigenschaften berechnet.

Abbildung 1:
Illustration des Verhältnisses von mathematischem Problem und Algorithmus, der zu dessen Lösung eingesetzt wird.

Algorithmische Entscheidungssysteme

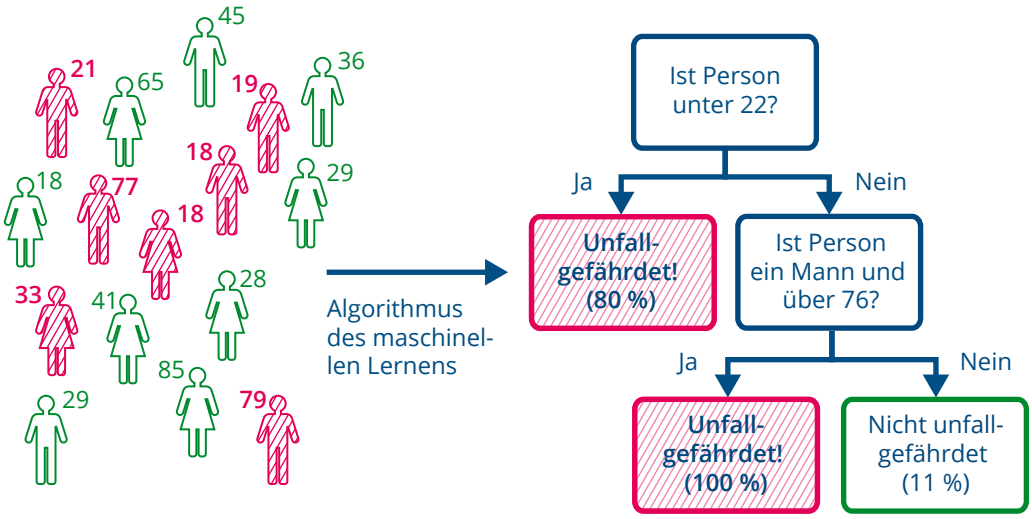


Abbildung 2:
Algorithmische Entscheidungssysteme bekommen Daten (Eingaben) und Informationen über ein zu lernendes Verhalten und leiten daraus Entscheidungsregeln ab. Im Beispiel sind Alter und Geschlecht von Autofahrern angegeben und die Information, ob sie in den letzten drei Jahren einen selbstverschuldeten Unfall hatten (rot: ja, grün: nein). Der Algorithmus findet Regeln, die die Daten möglichst gut so in zwei kleinere Mengen aufteilen, so dass in derselben Menge fast alle entweder einen Unfall hatten oder fast alle keinen. Diese Regeln können dazu verwendet werden, die Unfallgefährdung für neue Versicherungsnehmer abzuschätzen. Das Ergebnis ist stark von den verwendeten Daten abhängig.

Algorithmen des maschinellen Lernens

Informationen und Entscheidungsregeln

Es ist allerdings kein Algorithmus bekannt, der aus einer Reihe von Informationen, wie Alter, Geschlecht, finanzielle Situation u. a. berechnen könnte, wer künftig seinen Kredit zurückzahlt oder wer wieder kriminell wird. Es darf aber davon ausgegangen werden, dass es Regeln gibt, nach denen Menschen sich entscheiden, einen Kredit zurückzuzahlen oder wieder kriminell zu werden, oder dass in den Informationen über sie Hinweise auf Umstände enthalten sind, die ihr Verhalten bedingen. Diese Regeln können mit Hilfe von Algorithmen des sogenannten maschinellen Lernens aus den Informationen über die Personen und ihr Verhalten abgeleitet werden (s. Abbildung 2). Man gibt bspw. den Algorithmen des maschinellen Lernens Informationen über Alter und Geschlecht von Autofahrern und darüber, ob diese Personen in den letzten drei Jahren einen selbstverschuldeten Unfall hatten. Anschließend leitet der Algorithmus selbständig Entscheidungsregeln ab und teilt neu zu bewertende Personen fehlerfrei in zwei Kategorien ein: „unfallgefährdet“ oder „nicht unfallgefährdet“. Wie der Algorithmus die Entscheidungsregeln genau ableitet, hängt von der verwendeten Methode des maschinellen Lernens ab. Genauso wie der Umstand, ob sie für Menschen lesbar und einsichtig (wie im obigen Beispiel) sind oder nicht. Die abgeleiteten Regeln sind im Wesentlichen abhängig von der verwendeten Datengrundlage.

Warum handelt es sich hier um „Systeme“?

Daten und zwei Algorithmen

Algorithmische Entscheidungssysteme beinhalten Algorithmen an zwei Stellen: Der erste Algorithmus lernt auf Basis der Daten ein statistisches Modell. Das statistische Modell ist dann Grundlage für den (meist sehr einfachen) zweiten Algorithmus, der die eigentliche Entscheidung für eine neue Eingabe berechnet. In Abbildung 2 ist der eigentliche Entscheidungsalgorithmus in der baumartig organisierten Regelstruktur dargestellt und mündet

entweder in einem Kästchen, das den neuen Versicherungsnehmer als „unfallgefährdet“ oder „nicht unfallgefährdet“ beurteilt. Das Ergebnis eines algorithmischen Entscheidungssystems ist das Produkt aus der Interaktion von Daten und dem ersten Algorithmus. Man muss also immer das Gesamtsystem betrachten, bestehend aus Daten, dem ersten Algorithmus, der das Modell erlernt, und dem Modell, das dann die Grundlage für die Entscheidung bietet.

Klassische Algorithmen vs. algorithmische Entscheidungssysteme

Selbstlernende algorithmische Entscheidungssysteme

Dieser Ansatz, Entscheidungsregeln zu lernen, ist oft viel erfolgreicher als der Ansatz, einen klassischen Algorithmus zu entwickeln. Am Beispiel maschineller (d. h. algorithmischer) Übersetzungen von Texten lässt sich dies veranschaulichen: Hier wurden jahrzehntelang Algorithmen entwickelt, die dem alten Paradigma folgten: Gegeben ein Text in Sprache A, generiere eine Übersetzung des Textes in Sprache B. Aber selbst mit der Hilfe von vielen Linguisten konnten keine Algorithmen gefunden werden, um gute Übersetzungen zu generieren. Erst der radikal neue Ansatz, eine Vielzahl von Texten und ihre Übersetzungen als Datengrundlage zu nehmen und per Algorithmus daraus Entscheidungsregeln selbständig ableiten zu lassen, brachte den Durchbruch.

Allerdings können die Intransparenz der Entscheidungsregeln und die komplexen Interaktionen zwischen Daten, Algorithmen und sozialer Einbettung dazu führen, dass Entscheidungen einzelne Individuen oder die gesamte Gesellschaft schädigen. In solchen Fällen braucht es eine Überwachung ihrer Entscheidungsqualität und ihrer Einbettung in soziale Prozesse.

2. Welche Art von ADM-Systemen ist in der Transparenz- und Kontrolldiskussion relevant?

Erhöhter Kontrollbedarf

Algorithmische Entscheidungssysteme bedürfen dann einer speziellen Überwachung und Kontrolle, wenn sie Entscheidungen über Menschen treffen oder wenn ihre Entscheidungen einzelne Menschen oder die Gesellschaft als Ganzes betreffen. Die Kosten von Fehlentscheidungen trägt hier nicht nur der ADM-nutzende Akteur, sondern auch die Person, über die entschieden wird. Zu diesen sensiblen, der Überwachung und Kontrolle bedürfenden algorithmischen Entscheidungssystemen gehören folgende Beispiele (s. auch Lischka & Klingel, 2017):

1. Googles Suchmaschinenalgorithmus: Er lernt aus der Kombination von Informationen über eine Webseite und einen Nutzer und dem Nutzerverhalten, ob eine Webseite für eine bestimmte Suchanfrage relevant ist oder nicht.

2. Facebooks Newsfeed: Er lernt aus der Interaktion von Nutzern mit Inhalten, welche Informationen den Nutzern wichtig sind und sortiert nach den gelernten Entscheidungsregeln neue Inhalte.

3. Predictive Policing beruht auf Algorithmen, die aus bisherigen Straftaten ableiten, wann und wo welche neuen Straftaten zu erwarten sind. Diese ADM-Systeme werden derzeit auch in Deutschland entwickelt und erprobt.

4. Rückfälligkeitsvorhersagealgorithmen: In den USA werden Kriminelle mit Hilfe von ADM-Systemen in Risikogruppen etwa bezüglich weiterer Straftaten kategorisiert. Die Systeme lernen von Daten über Kriminelle, die in der Vergangenheit rückfällig wurden oder resozialisiert wurden. Sie leiten daraus Entscheidungsregeln ab, die in einer Gerichtsverhandlung oder bei der Vergabe von Resozialisierungsmaßnahmen verwendet werden können.

5. Terroristenidentifikation: Aus Dokumenten, die von Snowden veröffentlicht wurden, geht hervor, dass ADM-ähnliche Methoden eingesetzt werden sollten, um mögliche

terroristische Kuriere zu identifizieren, basierend auf Smartphone-Daten von 55 Millionen Einwohnerinnen und Einwohnern aus Pakistan und Afghanistan und wenigen bekannten Kurieren (Förtsch, 2015).

Grundsätzlich sind auch ADM-Systeme denkbar, die schlechte Entscheidungen über den Einsatz gesellschaftlicher Ressourcen treffen und damit einen Schaden verursachen, von dem der einsetzende Akteur nicht betroffen ist: beispielsweise ein ADM-System, das zwar den kostengünstigsten Transportweg für eine Reihe von Waren findet, der aber zugleich auch der umweltschädlichste ist. Noch ist nicht geklärt, inwieweit hier die ADM-Systeme selbst unter Aufsicht gestellt werden oder ob die Anreizstrukturen nicht eher grundsätzlich verändert werden müssen. An dieser Stelle sei angemerkt, dass falsche Anreizstrukturen mit Hilfe von Algorithmen optimal ausgebeutet werden und damit maximalen Schaden anrichten können.

Fehlurteile – direkte Betroffenheit von Menschen

Auf den Prüfstand müssen in jedem Fall ADM-Systeme, die durch Fehlurteile über Menschen diese direkt schädigen können. Um die daraus folgenden Transparenz- und Kontrollanforderungen zu erläutern, muss zunächst beschrieben werden, wie Fehlurteile entstehen können.

3. Fehlerquellen und Fehlurteile

Es gibt im Wesentlichen drei Mechanismen, die Fehlurteilen von ADM-Systemen zugrunde liegen.

1. Zufällige Faktoren: Einsatzgebiete von ADM-Systemen wie die Vorhersage künftiger Arbeits- oder Studienleistungen („people analytics“, s. z. B. Reindl & Krügl, 2017) unterliegen oft Zufälligkeiten und digital schwer zu erfassenden Dimensionen. Zufällig ist beispielsweise eine Erkrankung, die auch eine leistungsfähige Person einschränkt. Digital schwer messbar ist die Teamatmosphäre, die sich ebenfalls auf die Leistungsfähigkeit auswirkt. In diesen Fällen sind Fehlurteile fast zwangsläufig.

2. Zu kleine Datenmengen: Die Daten, von denen gelernt werden soll, sind zu inhomogen und ihre Zahl zu klein, um sie in homogenere Datenmengen aufzuteilen. So ist die automatische Identifikation von Terroristen auch deshalb so schwierig, weil es zu wenige Personen gibt, die aus denselben Gründen derselben Bewegung angehören und dort terroristisch aktiv werden. Auch hier sind Fehlurteile unvermeidbar.

3. Fehlerhafte Entwicklung oder fehlerhafter Einsatz von ADM-Systemen: Fehlurteile können aus einem fehlerhaft entwickelten ADM-System stammen oder einer Fehlinterpretation der von ihm gelieferten Resultate. Solche Fehlurteile können mithilfe strukturierter Entwicklungs- und Evaluationsprozesse minimiert werden.

Wie mit Fehlerquellen umgehen?

Die ersten beiden Faktoren, die zu Fehlurteilen führen, sind systemimmanent. Im Falle zu kleiner Datenmengen ist vom Einsatz von ADM-Systemen abzuraten. Nach Abwägung der Kosten und des Nutzens kann der Einsatz von ADM-Systemen in Situationen, die möglicherweise von Zufällen bestimmt sind, jedoch gerechtfertigt sein. Die auf fehlerhafter Entwicklung oder dem fehlerhaften Einsatz der Systeme beruhenden Fehlschlüsse können mittels Transparenz und Kontrolle reduziert werden. Nachfolgend werden einige wenige skizziert (vgl. ausführlich: Zweig, 2018).

1. Zu treffende Entscheidungen und mögliche Fehler in der Datenbasis:

   a. In der Eingabe gilt es, eine Auswahl von sinnvollen Informationen zu treffen. COMPAS, ein Rückfälligkeitsvorhersagealgorithmus, bekommt z. B. neben den Straftaten einer Person auch die Information, ob Verwandte des Kriminellen ebenfalls in Haft waren – in Deutschland undenkbar.

   b. Oftmals sind die Daten fehlerhaft oder werden bei der Kombination von mehreren Datenbanken nicht verlässlich der jeweiligen Person zugeordnet.

   c. Manchmal enthalten Daten schon Diskriminierendes, etwa wenn zu Bewerbungsverfahren Frauen oder Personen mit Migrationshintergrund weniger oft eingeladen worden sind, als sie unter den Bewerbern vertreten waren. Diese Diskriminierungen werden vom Algorithmus „mitgelernt". Wie genau „Diskriminierung" definiert werden soll und wie sie gemessen werden kann, ist auch eine gesellschaftliche Entscheidung (Kleinberg, Mullainathan & Raghavan, 2017; Zweig & Krafft, 2018).

2. Auch auf der Ebene des maschinellen Lernens müssen Entscheidungen getroffen werden und können Fehler gemacht werden. Nicht alle Algorithmen eignen sich für jede Frage.

3. Es muss eine Wahl getroffen werden, was genau zu lernen ist. Für algorithmische Entscheidungssysteme im Bereich von Social Media müssen die Designer eine vom Computer erfassbare Definition von „Relevanz" entwickeln. Dies stellt eine sogenannte „Operationalisierung" eines schwer greifbaren sozialen Konstrukts dar, die mehr oder weniger gut gelingen kann.

4. Die maschinelle Entscheidung muss – wenn sie von einem Menschen begutachtet wird – in einer Form kommuniziert werden, die alle oben genannten Entscheidungen und Unsicherheiten so fassbar macht, dass der menschliche Entscheider das Resultat bewerten kann.

5. Eine wichtige Erkenntnis ist, dass ein algorithmisches Entscheidungssystem für einen bestimmten sozialen Prozess entwickelt wird und nicht ohne weitergehende Prüfung in ähnlich scheinenden sozialen Prozessen eingesetzt werden darf (Zweig & Krafft, 2018). So wurde COMPAS, der Rückfälligkeitsvorhersagealgorithmus für Kriminelle, entwickelt, um Resozialisierungsmaßnahmen bestmöglich zu verteilen. Er wird aber inzwischen – unverändert – auch vor Gericht verwendet. Es konnte nachgewiesen werden, dass es sich um unterschiedliche soziale Prozesse handelt und dass die Qualität des Systems für die Anwendung vor Gericht nicht ausreichend ist. Hier müssen weitaus höhere Ansprüche an die Qualität gestellt und andere Aspekte der Entscheidungsqualität des Systems gemessen werden (Krafft, 2017).

ADM-Systeme sind folglich fehleranfällig, und oft ist es schwierig, bei Fehlentscheidungen die dafür Verantwortlichen zu identifizieren. Die Forderung nach einer Qualitätssicherung und mehr Transparenz ist daher berechtigt.

4. Wann ist es notwendig, ADM-Systeme zu überwachen?

Schädigungspotenziale

Die Auswirkungen von Fehlurteilen von ADM-Systemen, die Menschen bewerten, sind sehr unterschiedlich. Manche können einen hohen individuellen Schaden verursachen, werden aber nur auf wenige Personen angewendet. Andere verursachen einen geringen Schaden auf der Ebene von Individuen, aber einen hohen gesellschaftlichen Schaden.

Das Gesamtrisiko – also die Wahrscheinlichkeit für die Verwirklichung des Schadenspotenzials – ist aber ganz wesentlich noch von einer zweiten Dimension abhängig, nämlich der Leichtigkeit, mit der eine Zweitmeinung eingeholt werden kann bzw. Einspruch erhoben werden kann.

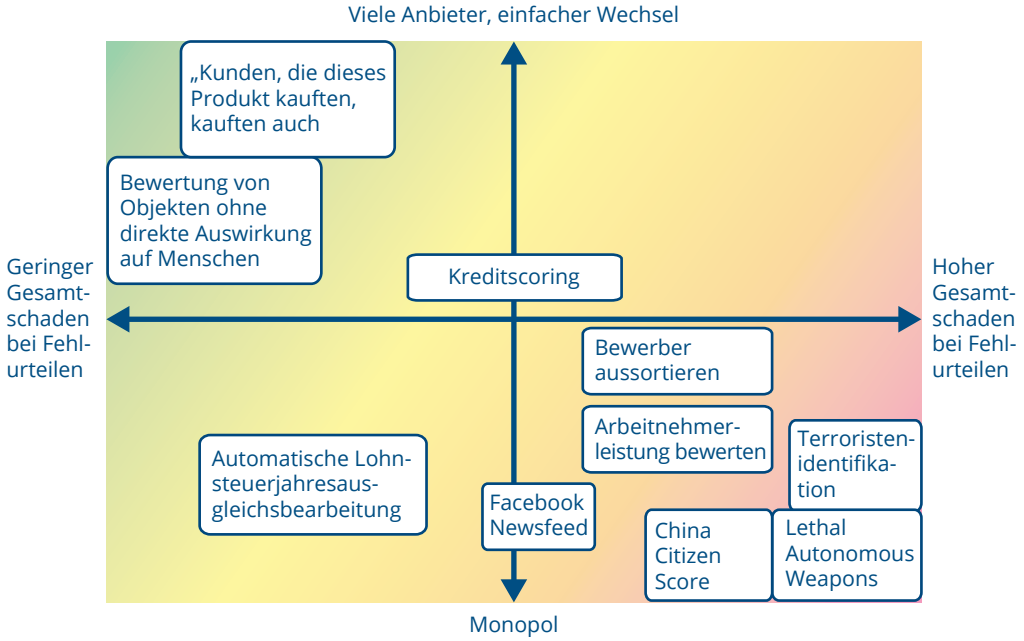


Abbildung 3: Risikomatrix mit einigen von der Autorin verorteten ADM-Systemen.

In Abhängigkeit von diesen beiden Dimensionen können verschiedene Anforderungen an Transparenz und graduell abgestufte Methoden der Kontrolle und Überwachung in Erwägung gezogen werden.

5. Mögliche Transparenzpflichten

Richtiger Umgang mit Transparenz

Häufig wird mehr „Transparenz" von algorithmischen Entscheidungssystemen gefordert. Der Begriff Transparenz wird jedoch sehr unterschiedlich interpretiert: Von der Forderung nach allgemeinen Angaben, welche Art von Informationen zum Lernen genutzt werden, über den Wunsch nach Erläuterung, wie Informationen jeweils zum Ergebnis beitragen, bis hin zur Forderung nach der Veröffentlichung des Codes. Insbesondere letztere erweist sich als wenig zielführend. Erstens ist der Softwarecode nicht lesbar wie ein Roman, und er kann beliebig kompliziert gemacht werden, wenn eine Firma ihre Firmengeheimnisse gefährdet sieht. Zweitens kann die Veröffentlichung schädlich sein: Die Veröffentlichung der Wirkweise des „PageRanks", dem damals grundlegenden Algorithmus der Suchmaschine von Google, führte z. B. dazu, dass die Webseitenbetreiber ihr Wissen ausnutzten, um die Suchmaschine zu manipulieren, ohne der Gesellschaft eine bessere Kontrolle zu ermöglichen. In bestimmten Fällen ist daher sogar „mehr Intransparenz" geboten, um Manipulationsrisiken zu minimieren.

Datentransparenz

Hinzu kommt, dass der Code allein nicht ausreicht, um die Wirkweise eines algorithmischen Entscheidungssystems nachzuvollziehen. Die Datengrundlage kann genauso wichtig sein, um Fehler zu erkennen. Unter Umständen muss auch transparent gemacht werden, wie genau die maschinelle Entscheidung in eine endgültige Entscheidung einfließt, etwa bei der Rückfälligkeitsvorhersage durch ADM-Systeme.

Die folgenden Stufen der Transparentmachung sind sinnvoll und könnten – je nach dem Schweregrad von Fehlurteilen durch ein ADM-System – durchgesetzt werden:

Schritte der Transparenzmachung

1. Transparente Darstellung der eingehenden Daten und des Kriteriums, mit dem das selbstständig lernende ADM-System trainiert wurde. Dieses Vorgehen erscheint für alle ADM-Systeme sinnvoll, die Menschen kategorisieren oder ihr künftiges Verhalten beurteilen.

2. Transparente Darstellung der groben Wirkungsweise des algorithmischen Entscheidungssystems und der von ihm gelernten Entscheidungsregeln in einer verständlichen Art und Weise. Hier ist es wichtig zu beachten, dass diese Transparenzmaßnahme eine große Einschränkung darstellt, da sie die wirkungsvollsten Methoden des maschinellen Lernens ausschließt, nämlich die, deren gelernte Entscheidungsregeln für Menschen nicht mehr nachvollziehbar sind.

3. Einblick in das ADM-System für einen ausgewählten Kreis von Experten.

Abhängig von potentiellen Schäden durch Fehlurteile können auch verschiedene Stufen der Überwachung angemessen sein.

6. Überwachung von ADM-Systemen

Sozio-informatisches Gesamtsystem

Aus den genannten Gründen sollten Kontrolle und Überwachung nicht allein am Algorithmus des maschinellen Lernens, dem eigentlichen Entscheidungsalgorithmus oder dem ADM-System an sich ansetzen. Im Fokus muss das sozio-informatische Gesamtsystem stehen, bestehend aus dem ADM-System und allen sozialen Akteuren, die es nutzen oder von dessen Entscheidungen betroffen sind (s. Abbildung 4). Alle qualitätssichernden Maßnahmen müssen die beiden Einzelsysteme und das Gesamtsystem umfassen.

Der wichtigste Punkt, um funktionale ADM-Systeme zu erhalten, ist ein qualitätsgesicherter Entwicklungsprozess für gesellschaftlich relevante ADM-Systeme. Ein solcher muss beispielsweise regeln, wer wann welche Entscheidung trifft, wie diese zu dokumentieren sind und welche Qualitätsmaßnahmen ergriffen wurden. Ein solch strukturierter Prozess fehlt bisher. Genauso wichtig ist ein qualitätsgesicherter Einbettungsprozess eines gesellschaftlich relevanten ADM-Systems in den sozialen Prozess, indem es verwendet werden soll. Dazu gehört unter anderem eine Einweisung für menschliche Entscheider, wie die Resultate der Maschine zu bewerten sind, und eine Prüfung, ob das ADM-System für den Prozess geeignet entwickelt wurde. Auch diesen strukturierten Prozess gibt es bisher nicht.

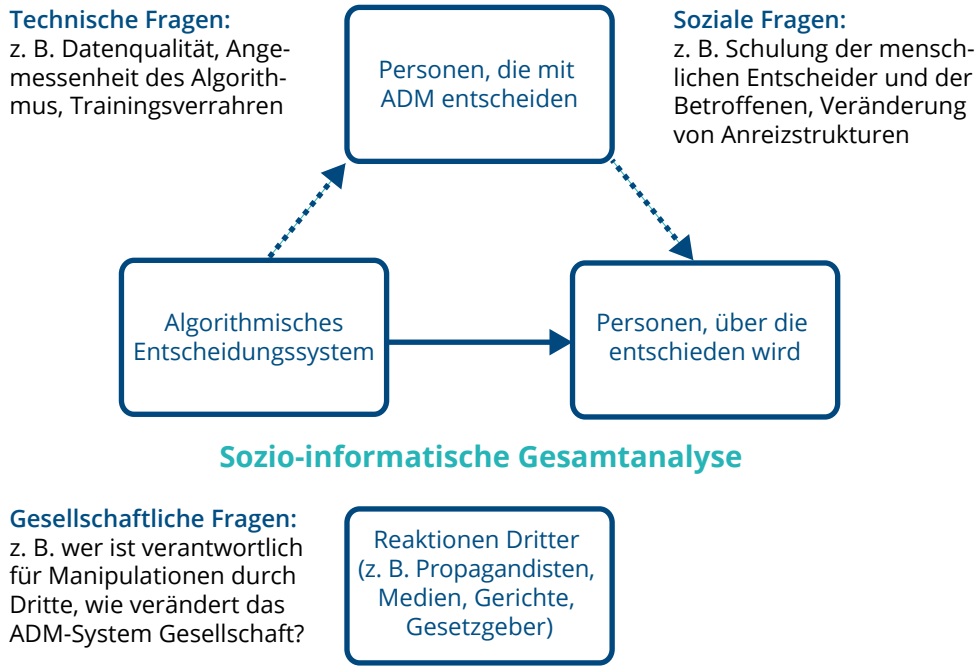


Abbildung 4:
ADM-Systeme verändern die Art und Weise, wie Entscheidungen in sozialen Prozessen gefällt werden. Je nachdem, wie transparent ihre Wirkungsweise kommuniziert wird, können sie auch gezielt genutzt werden. Sie können auch von Dritten eingeschränkt werden (z. B. Gerichten). Nicht zuletzt haben sie eine Auswirkung auf die Beurteilten, die darauf entsprechend reagieren könnten. Daher ist eine sinnvolle Evaluation der Auswirkungen des Einsatzes von ADM-Systemen nur möglich, wenn das sozio-informatische Gesamtsystem betrachtet wird.

Black-Box-Analysen

In vielen Fällen können die Auswirkungen von ADM-Systemen ohne eine genaue Kenntnis des darunterliegenden Wirkmechanismus überwacht werden mit sogenannten Black-Box-Analysen (Krafft & Zweig, 2018). Der wesentliche Bestandteil des Schadens liegt nämlich in den Anteilen der Fehlurteile des Systems. Wenn sich die beurteilten Personen darin beobachten lassen, wie sie sich weiterhin verhalten, kann dieser Anteil berechnet werden, ohne dass man weiß oder wissen muss, wie das System zu ihrem Urteil kam. Diese Beobachtung ist zum Beispiel für Suchmaschinen oder soziale Medien möglich. Diese treffen eine Auswahl von Nachrichteninhalten und können beobachten, ob Menschen tatsächlich mit dieser Auswahl interagieren oder nicht. In diesen Fällen kann die Fehlurteilsrate gemessen werden.

Diskriminierungspotenzial

Ein weiterer Gesichtspunkt, der sich gut durch eine Black-Box-Analyse überwachen lässt, ist die Frage nach Diskriminierung. Insbesondere bei algorithmischen Entscheidungssystemen im Bereich der (schulischen und beruflichen) Leistungsbewertung kann routinemäßig überprüft werden, ob Menschen nach ihrem Geschlecht, ihrer Herkunft, ihrem Alter oder weiteren Eigenschaften diskriminiert werden. Diese Überprüfung lässt sich leicht automatisieren, und sie sollte bei relevanten ADM-Systemen in Abhängigkeit von der Auswertung des möglichen Schadens gefordert werden.

Manchmal können auch ADM-spezifische Fragen mit Hilfe eines Black-Box-Ansatzes gelöst werden. Algorithmische Entscheidungssysteme stehen im Verdacht, durch starke Personalisierung Filterblasen zu verstärken oder überhaupt erst zu generieren (Eli Pariser 2011). Während dem Problem, ob sich Nutzer in einer Filterblase befinden oder nicht, methodisch

Personalisierung – Filterblasen

schwer zu begegnen ist, kann die Frage, wie stark personalisiert die Ergebnisse einer Suchmaschine sind, sehr gut automatisiert beantwortet werden. Ein solches Projekt hat das Algorithm Accountability Lab an der TU Kaiserslautern anlässlich der Bundestagswahl 2017 als proof of concept durchgeführt und festgestellt, dass die Suchergebnisse bei der freiwilligen Kohorte von Teilnehmern relativ wenig personalisiert waren (Krafft et al. 2017). Dieses Projekt zeigt, dass, auch wenn ein großer kollektiver Schaden z. B. durch Filterblasenbildung möglich ist, größere Eingriffe und Kontrollen nicht notwendig sind, solange ein Aspekt wie der Personalisierungsgrad als Grundlage einer möglichen Filterblase dauerhaft und kostengünstig überwacht werden kann.

In den meisten Fällen, in denen Maschinen Menschen beurteilen, wird den Menschen infolge der Beurteilung eine Handlungsoption entzogen, ohne dass sie ein Fehlurteil nachweisen können. Wer nicht zum Bewerbungsgespräch eingeladen wird, kann nicht unter Beweis stellen, dass er ein erfolgreicher Arbeitnehmer ist, und ein nicht vergebener Kredit kann nicht pünktlich zurückgezahlt werden. Hier liegt eine Grenze des Black-Box-Ansatzes.

Noch schwieriger wird es, wenn weder Betreiber noch Nutzer eines algorithmischen Entscheidungssystems angeben können, was die beste Entscheidung des Systems gewesen wäre. So etwa bei algorithmischen Entscheidungssystemen in Sozialen Medien, wo niemand weiß, was die beste Auswahl von Nachrichten gewesen wäre, aber auch bei Dating-Plattformen, bei denen keineswegs klar ist, wie man den besten Partner auswählt.

Aufgrund der geschilderten Schadensdimensionen durch Fehlurteile von ADM-Systemen müssen insbesondere ADM-Systeme der Geheimdienste, wie das genannte US-amerikanische System zur Identifikation terroristische Kuriere, intensiv kontrolliert und überwacht werden. Der individuelle Schaden für Unschuldig Identifizierte ist groß. Wenn Terroristen nicht identifiziert werden, ist der gesellschaftliche Schaden u. U. riesig. Die Zahl der beurteilten Personen ist hoch; die Überwachung findet vermutlich regelmäßig statt. Solche ADM-Systeme sollten einer hohen Transparenz (unter Umständen gegenüber einem akkreditierten, aber unabhängigen Expertenkreis) unterliegen und regelmäßig auf ihre gesamtgesellschaftlichen Auswirkungen hin überprüft werden. Ähnliches gilt für ADM-Systeme, die im Krieg zur Vermeidung von friendly fire eingesetzt werden, die Steuererklärungen nach Unregelmäßigkeiten durchsuchen oder die automatische Bewertung von Visa-Anträgen durchführen.

Die vorgestellten Transparenzpflichten und technischen Kontrollprozesse können nun verschiedenen Risikoklassen von ADM-Systemen mit gesellschaftlicher Relevanz zugeordnet werden.

7. Die fünf Risikoklassen von ADM-Systemen mit lernenden Komponenten

Ich schlage eine Unterteilung von gesellschaftlich relevanten ADM-Systemen in fünf verschiedene Klassen vor:

- ADM-Systeme der Klasse 0 müssen auf technischer Ebene nicht reguliert werden. Es gibt daher weder Transparenzpflichten noch die Notwendigkeit, diese ADM-Systeme ständig zu kontrollieren. Sollte ein Verdachtsfall auftreten, können post-hoc-Analysen angestrengt werden. Die Bewertung des Gesamtschadenspotenzials wird sich durch einen Verdachtsfall in den meisten Situationen verschlechtern und damit das System im Weiteren in einer höheren Klasse einer Regulation unterliegen. Zu dieser Klasse gehören beispielsweise Produktempfehlungssystem für Kleidung.

- Für ADM-Systeme der Klasse 1 empfehle ich eine ständige Überwachung des Systems durch eine Black-Box-Analyse, die keinen Zugriff auf den Code benötigt. Zu dieser Klasse gehört aus meiner Sicht Googles Suchmaschine, nachdem wir 2017 gute Belege dafür finden konnten (Krafft et al., 2018), dass der Personalisierungsgrad der dahinterliegenden Algorithmen so klein ist, dass eine Filterblase nicht zu befürchten ist. Eine solche Analyse sollte ständig durchgeführt werden, um bei einer Erhöhung des Personalisierungsgrades eine Neubewertung des möglichen Schadens vornehmen zu können.

- ADM-Systeme der Klasse 2 sind so kritisch, dass sie neben der ständigen Kontrolle auch mehrere Transparenzpflichten erfüllen müssen: Über die genaue Art der Eingabe, Qualität der Eingabedaten, das Qualitätskriterium und eventuelle Fairnesskriterien auch über die spezifische Einbettung des Systems in den sozialen Prozess, indem die finale Entscheidung getroffen wird. Zu dieser Klasse gehören beispielsweise automatische oder unterstützende ADM-Systeme, die eingehende Bewerbungen auf einen Job bewerten. Hier müsste beispielweise spezifiziert werden, welche der Daten aus den Bewerbungen verwendet wird, mit welcher Qualität die aus den Dokumenten ausgelesen werden können und mit welcher Maßgabe das ADM-System trainiert wurde. Es könnte beispielsweise die Anzahl der nicht erfolgreichen Jobinterviews minimieren oder aber die Anzahl der nachher erfolgreich eingestellten Personen maximieren – wenn unter „Erfolg“ verstanden wird, dass eine Person für mindestens zwei Jahre im Betrieb verbleibt, könnten sich z.B. Elternzeiten negativ auswirken. Nicht zuletzt müsste unter anderem transparent kommuniziert werden, ob die Entscheidung vollautomatisch getroffen wird oder nur der Vorbereitung der Entscheidung dient und welche Widerspruchsmöglichkeiten es gibt.

- ADM-Systeme, die in Klasse 3 verortet wurden, dürfen als lernende Komponente nur noch erklärende Modelle verwenden. Der oben gezeigte Entscheidungsbaum gilt den meistens als erklärend, da Menschen einigermaßen gut nachvollziehen können, wovon die Entscheidung abhängt. Momentan gehören neuronale Netze aus dem Bereich des Deep Learnings zu den Ansätzen, die als nicht-erklärend gelten. Hier ordnen wir beispielsweise Systeme ein, die innerhalb eines Betriebes versuchen, Arbeitnehmer nach zukünftigem Erfolg zu klassifizieren.

- Schlussendlich enthält die Klasse 4 solche Entscheidungssituationen, die aus meiner Sicht nicht durch algorithmische Entscheidungssysteme mit einer lernenden Komponente entschieden werden dürfen. Beispiele sind die automatische Tötung von vermeintlich erkannten, gesuchten Personen (lethal autonomous weapons), die flächendeckende Überwachung zur Bewertung bürgerlichen Verhaltens (Chinas Citizen Score) oder die Identifikation von Terroristen (Skynet).

Die fünf Risikoklassen sind in Abbildung 5 visualisiert.

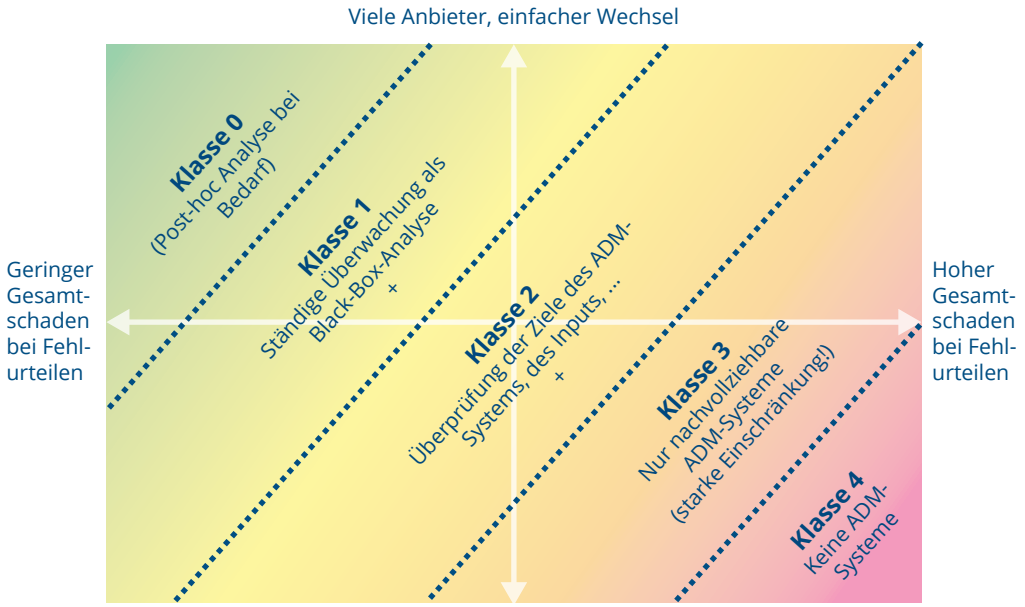


Abbildung 5:
Risikomatrix, auf der gesellschaftlich relevante ADM-Systeme in eine von fünf Risikoklassen eingruppiert werden können.

8. Politische Handlungsoptionen

Es ergibt sich eine Reihe von politischen Handlungsoptionen, die im Folgenden angesprochen werden.

8.1 Gesellschaftliche Diskussion über Transparenz und Kontrolle von algorithmischen Entscheidungssystemen

Deutschland ist eine Industrienation. Daher ist es wichtig, jegliche Regulationsbemühungen auf die Fälle zu beschränken, bei denen es auch tatsächlich notwendig ist. So müssen nicht alle algorithmischen Entscheidungssysteme (und schon gar nicht alle Algorithmen) durch den „Algorithmen-TÜV", es reicht aber auch nicht aus, sich auf die derzeit öffentlich stark wahrgenommenen algorithmischen Entscheidungssysteme der Sozialen Medien zu konzentrieren. In ihrem möglichen Schaden sind staatliche ADM-Systeme, aber auch die in den Firmen zunehmend verwendeten Bewerber- und Mitarbeiter-Bewertungssysteme mindestens ebenso wichtig – ganz zu schweigen von tödlichen, automatischen Waffensystemen (lethal autonomous weapons). In diesen Bereichen ist eine konstruktiv und effizient geführte Diskussion darüber notwendig, welche Art von ADM-Systemen in welchen sozialen Kontexten welcher Form der Kontrolle und Überwachung bedürfen. Einen Vorschlag zu einer grundsätzlichen Form der Klassifikation habe ich in dieser Studie vorgestellt.

8.2 Strukturierte Entwicklungs- und Einbettungsprozesse

Für die Entwicklung sozio-informatisch relevanter ADM-Systeme bedarf es eines qualitätsgesicherten Entwicklungsprozesses und eines strukturierten Einbettungsprozesses des ADM-Systems in das soziale System, das u. a. die Schulung der Nutzer und die Evaluation der Qualität der Entscheidungen umfasst. Hier könnten Bund und Länder bei der Beauftragung und dem Kauf von öffentlicher IT mit ADM-Systemen einen wichtigen Beitrag leisten, indem sie diesen strukturierten Prozesse erst entwickeln, dann den strukturierten Entwicklungsprozess bei den beauftragten Firmen fordern und den Einbettungsprozess der gekauften

Software in die eigenen sozialen Prozesse strukturiert begleiten und konstant die gesamtgesellschaftlichen Auswirkungen evaluieren.

8.3 Ausbildung, Berufsbilder und Berufsethiken

Es bedarf entsprechend qualifizierter Personen, die die notwendigen Fragen stellen und die die Qualität des Prozesses evaluieren können. Ein Studiengang, der solche Personen ausbildet, ist die (bisher in Deutschland einzigartige) Sozioinformatik an der TU Kaiserslautern.

Eine weitere Berufsgruppe, die hauptsächlich an der Entwicklung solcher ADM-Systeme beteiligt ist, sind data scientists. Das Berufsbild ist so neu, dass es bisher weder ein klares Curriculum gibt noch eine Berufsethik. Die meisten data scientists sind Quereinsteiger aus der Physik, Mathematik oder Informatik, mit großer Kompetenz in Statistik und Programmierung, denen jedoch in der Mehrzahl eine gesellschaftswissenschaftliche Ausbildung fehlt. Hier sind eine gezielte Förderung von Professuren im Bereich „data science“ und ein verbindliches Curriculum wichtig.

Zudem benötigen wir flächendeckend an allen Universitäten und Hochschulen, die Informatik anbieten, die früher oft anzutreffenden, inzwischen aber nahezu ausgestorbenen Professuren im Bereich „Informatik und Gesellschaft“.

Nicht zuletzt bedarf es auch einer gesonderten Berufsethik und vielleicht auch eines Akkreditierungsprozesses aller Personen, die die Inspektion von sozio-informatisch relevanten ADM-Systemen vornehmen sollen, um Bedenken der Firmen bezüglich ihrer Firmengeheimnisse Rechnung zu tragen.

9. Referenzen

Michael Förtsch, NSA-Überwachung: Skynet ist real und ziemlich unheimlich, Wired online, 11.5.2015, https://www.wired.de/collection/tech/das-nsa-programm-skynet-soll-terroristen-identifizieren (letzter Abruf: 12.11.18).

Jon Kleinberg, Sendhil Mullainathan & Manish Raghavan, Trade-Offs in the Fair Determination of Risk Scores, Proceedings of the 8th Innovations in Theoretical Computer Science Conference (ITCS’17), 2017, 43:1–43:23.

Tobias D. Krafft & Katharina A. Zweig, Wie Gesellschaft algorithmischen Entscheidungen auf den Zahn fühlen kann, in: (Un)berechenbar? Algorithmen und Automatisierung in Staat und Gesellschaft (Hrsg. Resa Mohabbat Kar, Basantha Thapa, Peter Parycek), Kompetenzzentrum Öffentliche IT, 2018, 471––492.

Tobias D. Krafft, Michael Gamer, Marcel Laessing & Katharina A. Zweig, Filterblase geplatzt? Kaum Raum für Personalisierung bei Google-Suchen zur Bundestagswahl, 1. Zwischenbericht des Datenspende-Projektes, 2017, https://algorithmwatch.org/wp-content/uploads/2017/09/1_Zwischenbericht__final.pdf (letzter Abruf: 12.11.18).

Tobias D. Krafft, Qualitätsmaße binärer Klassifikationen im Bereich kriminalprognostischer Instrumente der vierten Generation, Masterarbeit an der TU Kaiserslautern, 2017, https://arxiv.org/abs/1804.01557 (letzter Aufruf: 12.11.18).

Konrad Lischka und Anita Klingel, Wenn Maschinen Menschen bewerten – Internationale Fallbeispiele für Prozesse algorithmischer Entscheidungsfindung, Studie der Bertelsmann

Stiftung, Mai 2017, https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/ADM_Fallstudien.pdf (letzter Abruf: 12.11.18).

Viktor Mayer-Schönberger & Kenneth Cukier, Big Data: Die Revolution, die unser Leben verändern wird, Redline Verlag, München, 2013.

Eli Pariser, Filter Bubble: Wie wir im Internet entmündigt werden, Carl Hanser Verlag München, 2011.

Cornelia Reindl & Stefanie Krügl, People Analytics in der Praxis, Haufe-Lexware GmbH & Co. KG. Freiburg, 2017.

Katharina A. Zweig, Konrad Lischka & Dr. Sarah Fischer, Wo Maschinen irren können, Studie der Bertelsmann Stiftung, Februar 2018, https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/WoMaschinenIrrenKoennen.pdf (letzter Abruf: 12.11.18).

Katharina A. Zweig & Tobias D. Krafft, Fairness und Qualität algorithmischer Entscheidungen, in: (Un)berechenbar? Algorithmen und Automatisierung in Staat und Gesellschaft (Hrsg. Resa Mohabbat Kar, Basantha Thapa, Peter Parycek), Kompetenzzentrum Öffentliche IT, 2018, 204–227.

Impressum

Die Autorin

Prof. Dr. Katharina A. Zweig
Leiterin des Algorithm Accountability Labs des Fachbereichs Informatik an der TU Kaiserslautern. Dort ist sie auch Koordinatorin des bundesweit einzigartigen Studiengangs „Sozioinformatik“, der die Auswirkungen und Interaktion von Software und Individuum, Organisation und Gesellschaft modelliert, analysiert und, wo möglich, antizipiert. Prof. Zweig wurde mehrfach ausgezeichnet (z. B. ars legendi Fakultätenpreis in Informatik und Ingenieurswissenschaften 2017, Theodor-Heuss-Medaille 2018 als Mitgründerin der NGO AlgorithmWatch) und berät Landesmedienanstalten, die Kirchen und verschiedene Ministerien zu den gesellschaftlichen Auswirkungen der Digitalisierung. Seit September 2018 ist sie Mitglied der Enquete-Kommission „Künstliche Intelligenz“ des Bundestages.

Konrad-Adenauer-Stiftung e. V.

Dr. Norbert Arnold
Leiter des Teams Bildungs- und Wissenschaftspolitik
Hauptabteilung Politik und Beratung
T: +49(0)30 / 26 996-3504
norbert.arnold@kas.de

Postanschrift: Konrad-Adenauer-Stiftung, 10907 Berlin

Herausgeberin: Konrad-Adenauer-Stiftung e. V. 2018, Sankt Augustin/Berlin
Gestaltung: yellow too Pasiek Horntrich GbR
Satz: Janine Höhle, Konrad-Adenauer-Stiftung e.V.
Lektorat: Jenny Kahlert, Konrad-Adenauer-Stiftung e.V.

ISBN 978-3-95721-496-6

Der Text dieses Werkes ist lizenziert unter den Bedingungen von „Creative Commons Namensnennung-Weitergabe unter gleichen Bedingungen 4.0 international”, CC BY-SA 4.0 (abrufbar unter: https://creativecommons.org/licenses/by-sa/4.0/legalcode.de)

Bildvermerk Titelseite
© liuzishan, fotolia