Damit Maschinen den Menschen dienen

Lösungsansätze, um algorithmische Entscheidungen in den Dienst der Gesellschaft zu stellen

- Arbeitspapier -

BertelsmannStiftung

Damit Maschinen den Menschen dienen

Lösungsansätze, um algorithmische Prozesse in den Dienst der Gesellschaft zu stellen
- Arbeitspapier -

Julia Krüger
Konrad Lischka
im Auftrag der Bertelsmann Stiftung

Impressum
© Mai 2018
Bertelsmann Stiftung
Carl-Bertelsmann-Straße 256
33311 Gütersloh
www.bertelsmann-stiftung.de

Verantwortlich
Konrad Lischka, Ralph Müller-Eiselt

Autoren
Julia Krüger, Konrad Lischka

Lizenz
Dieses Arbeitspapier ist unter der Creative-Commons-Lizenz CC BY-SA 3.0 DE (Namensnennung – Weitergabe unter gleichen Bedingungen) lizenziert. Sie dürfen das Material vervielfältigen und weiterverbreiten, solange Sie angemessene Urheber- und Rechteangaben machen. Sie müssen angeben, ob Änderungen vorgenommen wurden. Wenn Sie das Material verändern, dürfen Sie Ihre Beiträge nur unter derselben Lizenz wie das Original verbreiten.

Titelbild: Hans/pixabay.com - CC0, Public Domain, https://creativecommons.org/publicdomain/zero/1.0/legalcode

DOI 10.11586/2018019 https://doi.org/10.11586/2018019

Inhalt

1 Vorwort 7

2 Worum es geht: Gesellschaftliche Anforderungen an algorithmische Entscheidungssysteme 9

2.1 Begriffliche Grundlagen 9

2.1.1 Vom Algorithmus zum Prozess algorithmischer Entscheidungsfindung 9

2.1.2 Lernende Systeme und schwache Künstliche Intelligenz 11

2.1.3 Entscheidungsunterstützung und automatisierte Entscheidungssysteme 13

2.1.4 Gesellschaftliche Teilhabe 14

2.2 Gesellschaftliche Anforderungen an algorithmische Prozesse 15

2.2.1 Gesellschaftliche Angemessenheit der Optimierungsziele 16

2.2.2 Umsetzung der Ziele in Systemen 17

2.2.3 Vielfalt der Systeme und Betreibermodelle 18

2.2.4 Schaffung übergreifender Rahmenbedingungen für teilhabeförderliche Systeme 19

3 Was zu berücksichtigen ist: Herausforderungen algorithmischer Systeme 21

3.1 Einsatzfeld: Sind teilhaberelevante Fragen berührt? 21

3.2 Zielsetzung und Evaluation: Wer definiert und kontrolliert Erfolg – und wie? 23

3.3 Dynamik und Komplexität: Wie entwickelt sich das Entscheidungssystem? 23

3.4 Automatisierung: Wie eigenständig agiert das Entscheidungssystem? 25

3.5 Sicherheit: Wie gut ist das Entscheidungssystem gegenüber Manipulation geschützt? 26

3.6 Zwischenfazit 27

4 Was man tun kann: Panorama der Lösungsvorschläge 29

4.1 Zielsetzung algorithmischer Systeme auf gesellschaftliche Angemessenheit prüfen 29

4.1.1 Interessen, Stakeholder und Optimierungsziele dokumentieren 30

4.1.2 Betroffene über den ADM-Einsatz informieren 31

4.1.3 Die erwarteten Folgen und Auswirkungen reflektieren und dokumentieren 34
4.1.4 Partizipation aller Stakeholder bei Entwicklung und Anwendung sichern 35
4.1.5 Professionsethik etablieren 36
4.2 Umsetzung von Zielen in Systemen 39
4.2.1 Methoden zur Umsetzungsprüfung entwickeln 38
4.2.2 Qualität der Datenbasis verbessern und dokumentieren 42
4.2.3 Überprüfbarkeit algorithmischer Systeme gesetzlich ermöglichen und sichern 44
4.2.4 Widerspruchsmöglichkeiten bei algorithmischen Prozessen institutionalisieren 46
4.2.5 Öffentliche Aufsicht über algorithmische Systeme entwickeln 48
4.2.6 Zivilgesellschaftliches Engagement fördern 49
4.3 Diversität schaffen 49
4.3.1 Vielfalt durch zugängliche Trainingsdatensätze stärken 51
4.3.2 Staatliche Nachfrage nach algorithmischen Systemen zur Vielfaltssicherung nutzen 53
4.3.3 Gemeinwohlorientierte Entwicklung algorithmischer Prozesse fördern 55
4.4 Übergreifende Rahmenbedingungen für teilhabeförderlichen ADM-Einsatz schaffen 56
4.4.1 Gesetzlichen Rahmen auf Anpassungsbedarf prüfen 56
4.4.2 Staatliche Regulierungskompetenz stärken 58
4.4.3 Individuelle Sensibilisierung und Kompetenz 59
5 Zusammenfassung und Fazit: Was nun zu tun ist 63
5.1 Die Ziele und Mittel: Gesellschaftliche Angemessenheit prüfen 64
5.2 Die Wirkung: Umsetzung von Zielen in algorithmischen Systemen prüfen 65
5.3 Die Vielfalt: Diversität algorithmischer Systeme und Prozesse sichern 68
5.4 Der Rahmen: Recht, staatliches Können, individuelle Kompetenzen 69
5.5 Jetzt handeln! 69
6 Literatur 71
7 Über die Autoren 83

8 Impulse Algorithmenethik 84

1 Vorwort

Welche weiterführenden Schulen dürfen Kinder besuchen? Wo fährt die Polizei Streife? Wessen Steuererklärung wird von Menschen, welche ausschließlich von Software bearbeitet? Welche Passanten am Bahnhof gelten als verdächtig und welche Angeklagten vor Gericht als besonderes Risiko? Bei solchen Entscheidungen und Prognosen setzen Staat und Unternehmen weltweit auf algorithmische Systeme. Algorithmische Systeme wirken in immer mehr Lebensbereichen und haben Einfluss auf das Leben von immer mehr Menschen. Automatisierung und Algorithmisierung erreichen eine neue Qualität, weil Informationstechnik allgegenwärtig ist, mehr Daten digital erfasst und Analyseergebnisse leichter umgesetzt werden können.

Bei der Debatte über den Einsatz solcher Systeme kommt es leicht zu gedanklichen Kurzschlüssen wie diesen: Die Algorithmen bestimmen über uns. Selbstlernende Systeme lernen automatisiert, worüber sie entscheiden. Eine solche Argumentation verschleiert Verantwortlichkeiten und verkennt, wie algorithmische Systeme tatsächlich entstehen und funktionieren.

In der Praxis definieren Menschen das von algorithmischen Prozessen zu bearbeitende Problem und den dabei anzustrebenden Zustand. Zwei Beispiele für solche Zielformulierungen: Gewinne ein Spiel wie Schach oder Go und befolge dabei folgende Spielregeln. Oder: Verteile möglichst viele Schüler ihren Präferenzen und dem Bedarf entsprechend auf die verfügbaren Plätze an weiterführenden Schulen – ohne etwas an der Menge verfügbarer Plätze und etwaigem Mangel zu ändern. Der zweite Fall zeigt, dass Menschen oft schon bei der Definition des zu lösenden Problems und somit vor der Konstruktion eines algorithmischen Systems Entscheidungen über die praktische wie auch die gesellschaftliche Wirkung fällen. Konstruiere ich ein System, um das Angebot zu verbessern? Oder entwickle ich wie hier im Beispiel ein System, das im Rahmen des bestehenden Angebots die vielleicht zu knappen Ressourcen besser verteilen soll?

Welche Optimierungsziele gesellschaftlich sinnvoll sind, lässt sich nicht allgemein für alle algorithmischen Systeme festlegen. Definition und Priorisierung gesellschaftlicher Ziele sind ein dynamischer Prozess. Jedes neue algorithmische System treibt diese Aushandlung weiter. Daraus folgt: Wer solche Systeme mit Wirkung auf Teilhabechancen beauftragt, entwickelt und einsetzt, muss auch einen angemessen breiten gesellschaftliche Diskurs darüber ermöglichen und dabei potenziell betroffene Menschen einbeziehen. Denn bei allem Fortschritt in einigen Einsatzgebieten Künstlicher Intelligenz (KI) ist diese menschliche Fähigkeit unerreicht: Ziele gemeinsam mit anderen festzulegen und das kollektive Verständnis zu etablieren, dass diese Ziele erstrebenswert sind.

Das vorliegende Arbeitspapier behandelt genau diese Schnittstelle von Technik und Gesellschaft. Es skizziert Lösungsideen für Politik, Wissenschaft, Zivilgesellschaft (z. B. Verbraucherschutzorganisationen, andere Nichtregierungsorganisationen, Aktivisten), Wirtschaft und Entwickler. Es bietet Orientierung für alle Akteure und einen Baukasten für Entscheider zur Gestaltung algorithmischer Systeme, um solche Systeme in den Dienst der Gesellschaft zu stellen und Teilhabe aller zu stärken statt Menschen oder Gruppen vom gesellschaftlichen Fortschritt auszuschließen. Das Papier zeigt das Spektrum der bisher diskutierten Lösungsansätze auf und systematisiert den Überblick. Dazu beantwortet es diese Kernfragen: Welche Herausforderungen sind bei der gesellschaftlichen Gestaltung algorithmischer Systeme erkennbar (Kapitel 2 und 3)? Welche Optionen gibt es, um diesen Herausforderungen zu begegnen (Kapitel 4 und 5)?

Unser Dank gilt Dr. Ulf Buermeyer, Dr. Andreas Dewes, Prof. Dr.-Ing. Florian Gallwitz, Lorena Jaume-Palasí, Dr. Nicola Jentzsch und Philipp Otto für ihre kritischen Prüfungen und wertvollen Anregungen.

Wir veröffentlichen das Arbeitspapier unter einer freien Lizenz (CC BY-SA 3.0 DE), um einen Beitrag zu einem sich schnell entwickelnden Feld zu geben, auf dem auch andere aufbauen können. Über Erweiterungen, Verbesserungen, weiterführende Analysen und natürlich auch konstruktive Kritik freuen wir uns sehr. Die Bertelsmann Stiftung will das Entwickeln, Konkretisieren und Erproben ausgewählter Lösungsansätze fördern – wir freuen uns über Interessenten.

Die Analyse ist Teil des Projekts „Ethik der Algorithmen“, in dem sich die Bertelsmann Stiftung näher mit den gesellschaftlichen Auswirkungen algorithmischer Entscheidungssysteme beschäftigt. Bislang sind eine Reihe von Impulsen erschienen: eine Sammlung internationaler Fallbeispiele (Lischka und Klingel 2017), eine Untersuchung des Wirkungspotenzials algorithmischer Entscheidungsfindung auf Teilhabe (Vieth und Wagner 2017), eine Analyse des Einflusses algorithmischer Prozesse auf den gesellschaftlichen Diskurs (Lischka und Stöcker 2017) sowie ein Arbeitspapier zu Fehlerquellen in Prozessen algorithmischer Entscheidungsfindung (Zweig 2018) und eine Analyse der Datenschutzgrundverordnung.

Ralph Müller-Eiselt
Senior Expert
Taskforce Digitalisierung
Bertelsmann Stiftung

Konrad Lischka
Projektleiter
Ethik der Algorithmen
Bertelsmann Stiftung

2 Worum es geht: Gesellschaftliche Anforderungen an algorithmische Entscheidungssysteme

In der Debatte über den Einsatz algorithmischer Entscheidungssysteme geht vieles durcheinander: Was kann Software, was kann sie nicht? Wie sollten Menschen mit algorithmischen Prognosen umgehen? Welche Optimierungsziele sollten öffentliche und private Akteure verfolgen und welchen Unterschied macht es, ob sie dafür algorithmische oder andere Entscheidungssysteme nutzen? Das folgende Kapitel versucht, relevante Aspekte dieser Diskussion zu definieren. Nach einer Einführung in begriffliche Grundlagen beantwortet es strukturiert die Fragen: Was sind die Herausforderungen algorithmischer Entscheidungsprozesse? Welche gesellschaftlichen Anforderungen ergeben sich daraus für ihre Anwendung und welche Lösungsansätze sind derzeit bekannt?

2.1 Begriffliche Grundlagen

Algorithmen, so sagt man vereinfacht, steuern heute vielfach Entscheidungsprozesse (wie z. B. bei der Gesichtserkennung im Polizeieinsatz) oder erarbeiten Analysen als Grundlage für menschliche Entscheidungen (wie z. B. bei der Einsatzplanung von Polizeistreifen). Damit ist ein komplexes Zusammenspiel von technischer und sozialer Organisation gemeint, das unten weiter differenziert wird. Die daraus hervorgehenden algorithmischen Entscheidungssysteme und -prozesse werden in der öffentlichen Debatte häufig im Zusammenhang mit der Entwicklung sogenannter **Künstlicher Intelligenz (KI)** diskutiert. Dies ist insofern korrekt, als dass es bei den algorithmischen Entscheidungen im aktuellen Diskurs insbesondere um die Automatisierung solcher Entscheidungen geht, die bislang nur Menschen treffen konnten und die ohne die Entwicklung Künstlicher Intelligenz nicht möglich wäre (Ramge 2018). Allerdings erfordert eine Debatte über Herausforderungen und Lösungsansätze im Bereich algorithmischer Entscheidungsfindung mehr begriffliche Klarheit. Daher folgt zunächst eine Definition der wichtigsten Begrifflichkeiten, die wesentliche Bezüge zu Künstlicher Intelligenz integrieren.

2.1.1 Vom Algorithmus zum Prozess algorithmischer Entscheidungsfindung

Die Grundlage algorithmischer Entscheidungsfindung ist der **Algorithmus**: Er bezeichnet im allgemeinen Sprachgebrauch eindeutige Handlungsvorschriften zum Lösen eines vorab definierten Problems, die in Mathematik und Informatik von größter Bedeutung sind. Die in Programmiersprache (Code) formulierten Regeln geben hier einen Plan vor, nach welchem Eingabedaten zu einem bestimmten Zweck verarbeitet und in Ausgabedaten umgewandelt werden. Die Resultate sind die Grundlage für eine algorithmische Entscheidung.

Technisch umgesetzte Algorithmen sind

> *„(…) informatische Werkzeuge, um mathematische Probleme automatisiert zu lösen. Sie berechnen zuverlässig eine Lösung für ein Problem, wenn sie die dafür nötigen Informationen bekommen, den sogenannten Input. Das mathematische Problem definiert, welche Eigenschaften der dazugehörige Output, also das Resultat der Berechnung, haben soll“* (Zweig 2018: 9).

Ein Algorithmus beschreibt einen Lösungsweg, der im Regelfall erst durch die Implementierung von Software in Computern wirksam wird. Solche **algorithmischen Systeme (Algorithmic-Decision-Making-Systeme, ADM-Systeme)** dienen der Lösung eines spezifischen Problems (vgl. Abbildung 1). Die entwickelten Entscheidungssysteme umfassen als Software unter anderem:

- Ein- und Ausgabedaten,
- eine Operationalisierung des zu lösenden Problems,
- Modelle für die Anwendung der Algorithmen zur Entscheidungsfindung.

Kurz: „*A fully configured algorithm will incorporate the abstract mathematical structure that has been implemented into a system for analysis of tasks in a particular analytic domain*“ (Mittelstadt et al. 2016: 7).

Wenn ein algorithmisches System zum Beispiel die Relevanz von Mitteilungen in einem sozialen Netzwerk bestimmen soll, muss Relevanz operationalisiert und gemessen werden, etwa über die Anzahl von positiven Reaktionen auf den zu bewertenden Beitrag. Diese Daten müssen erfasst und systematisch ausgewertet werden. Das Optimierungsziel des Prozesses muss messbar gemacht werden, zum Beispiel so: Die Relevanz von angezeigten Inhalten bemisst sich an der Menge der Kommentare, Favorisierungen und vergleichbaren Reaktionen.

Algorithmische Systeme können in vielfältigen Bereichen zum Einsatz kommen, die in ihrer Komplexität stark variieren. Wir kennen seit Langem etwa Fahrkartenautomaten und Fertigungsroboter. In der aktuellen Debatte geht es vor allem um die Automatisierung von Entscheidungen, die bislang nur Menschen treffen konnten, etwa LKW-Fahrer, Sachbearbeiter oder Ärzte[1]. Die Einbettung von algorithmischen Systemen in Anwendungsbereichen der Wissensarbeit ist komplex. Daher verwenden wir den Begriff **Prozess algorithmischer Entscheidungsfindung (ADM-Prozess)**, um die Gesamtheit aus algorithmischem System und seiner gesellschaftlichen Einbettung zu beschreiben: Das System wird in bürokratische oder organisatorische Verfahren eingebettet, Menschen entscheiden in diesem Prozess, wie sie den Output des algorithmischen Systems umsetzen, andere evaluieren dieses System und so weiter.

Damit wird das Zusammenwirken von Mensch und Maschine deutlich: Das Definieren der Ziele, auf die ein System hin optimiert wird, ist beispielsweise kein rein technischer Prozess. Ebensowenig sind es die Datenauswahl, die Messbarmachung sozialer Konstrukte wie Nachrichtenrelevanz oder die Interpretation von Ergebnissen:

> *„Algorithmische Entscheidungsfindung basiert immer auf bestimmten Werten und Normen. Daher kann nie nur der Algorithmus ‚an sich' untersucht werden, sondern es muss immer auch die Einbettung in einen sozialen Kontext berücksichtigt werden"* (Vieth und Wagner 2017: 11).

Wenn beispielsweise ein algorithmisches System daraufhin optimiert wird, an der Onlinenutzung labile Personen in Krisensituationen zu erkennen, um ihnen zielgerichtet bestimmte Werbung oder Unterstützungsangebote anzuzeigen, ist das womöglich technisch zuverlässig umsetzbar. Unabhängig davon ist zu entscheiden, ob dieser Einsatz gesellschaftlich angemessen ist.

Die gesellschaftliche Einbettung eines algorithmischen Systems umfasst auch die Verfahren zur Evaluation eines Systems, die Beschwerde- und Widerspruchsmöglichkeiten für die Betroffenen und die Korrekturpraxis: Reichen die personellen Ressourcen? Sind Mitarbeiter im Umgang mit dem algorithmischen System geschult? Können sie Prognosen sachgerecht einordnen und anwenden? Erlauben die jeweiligen Abhängigkeitsstrukturen vor Ort Mitarbeitern eine freie Entscheidung, sich über die Prognosen eines algorithmischen Systems hinwegzusetzen? Beispiele wie die Zweckentfremdung der „Strategic Subject List" in Chicago zeigen, dass Fehler bei der Implementierung eines Systems in einer Behörde zum Misserfolg des Gesamtprozesses führen können – unabhängig von der Qualität des genutzten algorithmischen Systems: Hierbei handelt es sich um eine Liste, mithilfe derer das individuelle Risiko für eine Verwicklung in Straftaten berechnet werden kann. Sie wurde zur Prävention entwickelt (City of Chicago 2017). Im Ergebnis diente sie allerdings der Ermittlung von Verbrechen (Kunichoff und Sier 2017).

Der Begriff **Prozess algorithmischer Entscheidungsfindung (ADM-Prozess)** meint also ein komplexes Arrangement von technischer und menschlicher Entscheidungsfindung. Als Abkürzung nutzen wir **ADM-Prozess,** nach dem im Englischen üblichen Begriff **Algorithmic Decision-Making** (Jaume-Palasí und Spielkamp 2017; Tene und Polonetsky 2017; Wachter, Mittelstadt und Russell 2017). ADM-Prozesse sind derzeit vor allem in der Wirtschaft und in geringem Ausmaß bei staatlichem Handeln im Einsatz: in der Wirtschaft zum Beispiel bei der Bewerberauswahl, dem Erstellen von Konsumentenprofilen, der Marktsegmentierung oder der Personalisierung von Angeboten. Staatliche Akteure nutzen ADM-Prozesse in der Polizeiarbeit, zum Beispiel bei der Einsatzplanung von Polizeistreifen, der automatisierten Gesichts- und Verhaltenserkennung, wie im Pilotprojekt am Bahnhof Berlin Südkreuz,

[1] Aus Gründen der Einfachheit und besseren Lesbarkeit verwendet diese Publikation vorwiegend die männliche Sprachfom. Es sind jedoch jeweils beide Geschlechter gemeint.

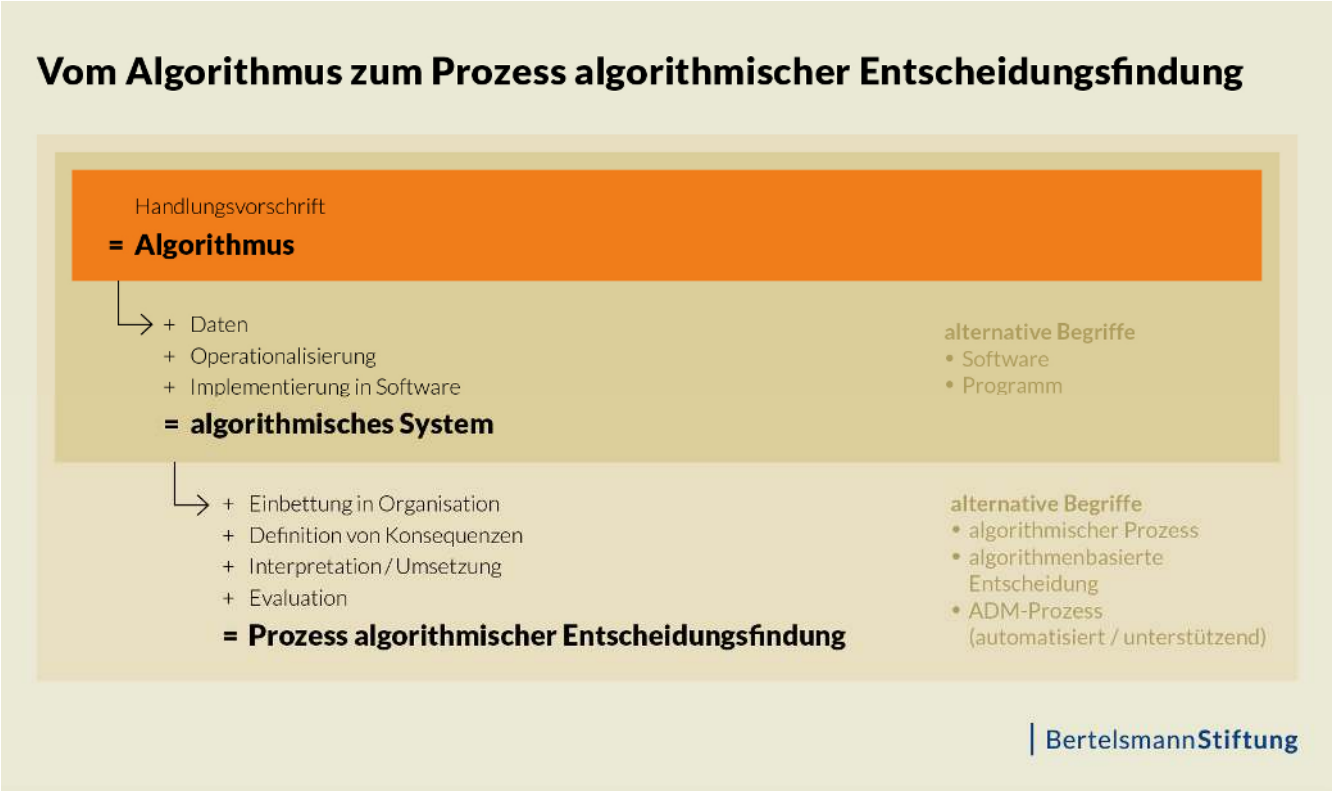
oder der Akzenterkennung von Antragsstellern in Asylverfahren (Bundesamt für Migration und Flüchtlinge 2017). Mögliche Anwendungsszenarien sind vielfällig.

Wir sind heute mit sehr unterschiedlichen algorithmischen Entscheidungssystemen konfrontiert, die jeweils ein spezifisches Problem mit menschenähnlicher oder sogar Menschen übertreffender Leistung lösen können. Zum Beispiel: Go spielen, Personen auf Fotos identifizieren oder gesprochenes Englisch verschriftlichen. Für diese Vielfalt hat der Robotiker Hans Moravec das Bild einer Landschaft menschlicher Kompetenzen entwickelt:

> *„Imagine a ‚landscape of human competence', having lowlands with labels like ‚arithmetic' and ‚rote memorization', foothills like ‚theorem proving' and ‚chess playing', and high mountain peaks labeled ‚locomotion', ‚hand−eye coordination' and ‚social interaction'. We all live in the solid mountaintops, but it takes great effort to reach the rest of the terrain, and only a few of us work each patch. Advancing computer performance is like water slowly flooding the landscape. A half century ago it began to drown the lowlands, driving out human calculators and record clerks, but leaving most of us dry. Now the flood has reached the foothills, and our outposts there are contemplating retreat. We feel safe on our peaks, but, at the present rate, those too will be submerged within another half century"* (Moravec 1998: 20).

Heute, 20 Jahre später, sind einige der Landstriche überflutet – etwa die Go-Hochebene. Welche Anwendungen sich daraus ergeben, welche algorithmischen Entscheidungssysteme in ADM-Prozesse überführt werden, welche menschlichen Entscheidungen durch Maschinen ersetzt werden, bleibt abzuwarten und durch Diskurs und Regulierung zu gestalten.

Abbildung 1:



(Quelle: eigene Darstellung)

2.1.2 Lernende Systeme und schwache Künstliche Intelligenz

Jenseits der Komplexität ihrer gesellschaftlichen Einbettung können ADM-Prozesse auf lernenden und nicht lernenden Algorithmen bzw. algorithmischen Systemen beruhen. Die Unterscheidung ist wesentlich, denn sie hat eine große Bedeutung dafür, ob und wie sich ADM-Prozesse prüfen und kontrollieren lassen (vgl. Kapitel 4).

Algorithmische Systeme, die mit einer größeren Automatisierung Modelle zur Problemlösung erstellen, bezeichnen wir als **lernende Systeme** oder synonym als **schwache Künstliche Intelligenz**. Damit meinen wir aus Daten lernende Software, die „in Verbindung mit steuerungsfähiger Hardware den Dreischritt Erkennen, Erkenntnis und Umsetzung in eine Handlung" ermöglicht (Ramge 2018: 14).

Dabei kommen vor allem **Formen des maschinellen Lernens** zum Einsatz: Algorithmen suchen in Daten Muster, speichern diese und wenden sie auf neuen Input an.[2] Dabei wird der Algorithmus ziel- und nutzenorientiert fortentwickelt, zum Zwecke einer effektiven sowie Zeit und Speicherkapazität schonenden Datenanalyse und -verarbeitung (Russel und Norvig 2012). Ziel- und Nutzendefinition erfolgt durch den Menschen und stellt den wesentlichen Unterschied zu Science-Fiktion-Dystopien sogenannter **starker Künstlicher Intelligenz** dar. Zentrale Grundlagen für die Umsetzung des maschinellen Lernprozesses sind **neuronale Netze**[3] und verschiedene Suchalgorithmen, die in unterschiedlichen Lernstilen fortentwickelt werden können:

> *„Algorithmen suchen in Daten nach statistisch auffälligen Mustern. Diese Informationen speichern sie in verschiedenen Arten von Strukturen, z. B. einer mathematischen Formel, in Entscheidungsbäumen oder einem neuronalen Netz. Diese Struktur wird ‚Modell' genannt"* (Zweig 2018: 17).

Zu den wichtigsten Lernstilen gehören:

- **überwachtes Lernen („supervised learning")**: Das System wird mit Daten trainiert, deren korrekte Klassenzugehörigkeit zu Beginn des Trainings bekannt ist. Diese Vorgabe („ground truth") kann auf menschlichem (Experten-)Wissen beruhen, zum Beispiel hinsichtlich der Einschätzung, ob es sich bei einer abgebildeten Person um einen Mann oder eine Frau handelt. Bei prognostisch eingesetzten algorithmischen Systemen kann sich die Klassenzugehörigkeit von ausreichend alten Datensätzen auch ganz ohne menschliches Zutun ergeben, etwa bei der Fragestellung, ob das Profil eines Studienbewerbers den Profilen erfolgreicher Absolventen ähnelt (Gallwitz 2018).
- **unüberwachtes Lernen („unsupervised learning")**: Die Trainingsdaten sind nicht vorab klassifiziert, das System erarbeitet selbst basierend auf Mustern eine Klassifikation (nach Böttcher, Klemm und Velten 2017: 8). Ein Beispiel für ein solches Lernverfahren wäre die Videokameraüberwachung in einem Einkaufszentrum, welche selbstständig Personen wiedererkennt und dadurch Prognosen über die Besuchsfrequenz von Kunden erstellt. Stimmen mehrere Abbildungen von Personen in bestimmten Merkmalen in ausreichendem Maße überein, dann geht das System davon aus, dass es sich um dieselbe Person handeln muss.
- **teilüberwachtes Lernen („semi-supervised learning")**: Es handelt sich hierbei um eine Mischform von überwachtem und unüberwachtem Lernen. Nur ein kleiner Teil der Daten ist vorab klassifiziert, aber die Art und Anzahl der Klassen wird wie beim überwachten Lernen vorgegeben.
- **bestärkendes Lernen („reinforcement learning")**: Die menschlichen Entwickler definieren Erfolg und belohnen Handlungen des Systems, die Erfolge erzielen. Solche Verfahren sind heute vor allem dort erfolgreich, wo algorithmische Systeme in geschlossenen (häufig virtuellen oder simulierten) Welten autonom agieren sollen, etwa bei der automatisierten Bewältigung von Atari-Videospielklassikern.

Die Aufzählung macht einen wesentlichen Unterschied zu **regelbasierten, nicht lernenden algorithmischen Systemen** deutlich: Künstlich intelligente Systeme enthalten Feedbackschleifen. Sie messen die Auswirkungen

---

[2] Die Einschränkung der hier betrachteten Künstlichen Intelligenz auf maschinelles Lernen ergibt sich daraus, dass diese Technologie derzeit Vorreiter in der Entwicklung ist (Ramge 2018). Eine annähernde Auflistung aller Technologien, die zur Künstlichen Intelligenz gezählt werden, würde den Rahmen dieses Arbeitspapieres sprengen.

[3] Neuronale Netze bezeichnen eine beliebige Anzahl miteinander verbundener Neuronen und bilden so einen Teil des Nervensystems von Lebewesen. In der Entwicklung der Künstlichen Intelligenz nehmen neuronale Netze eine besondere Stellung ein: Durch Nachbildung mittels Daten und Algorithmen werden hier digitale Problemlösungsarchitekturen geschaffen, die parallele, nicht lineare und komplexe Informationsverarbeitung ermöglichen (für eine detaillierte Beschreibung siehe Russel und Norvig 2012).

ihrer Entscheidungen und beziehen sie in die Ergebnisse folgender Entscheidungsprozesse ein (Ramge 2018). Die Verbesserung der Ergebnisse bzw. die in das System eingebaute Korrektur kann verschiedene Automatisierungsgrade annehmen, die hier nicht weiterverfolgt werden. Das Wichtige ist: Die Entwickler definieren nicht jeden einzelnen Lösungsschritt vorab, die Modellbildung erfolgt teilautomatisiert und sukzessiv. Dennoch bleibt die Evaluation des gesamten ADM-Prozesses von elementarer Bedeutung.

Selbstlernende algorithmische Systeme befinden sich derzeit in einer rasanten Entwicklung. Viele bekannte ADM-Prozesse basieren aber auf **nicht lernenden algorithmischen Systemen** und sind deshalb ebenfalls Gegenstand dieses Arbeitspapiers. Wir verstehen sie als Systeme, deren Modelle zur Problemlösung Schritt für Schritt von menschlichen Entwicklern geschaffen werden. Vorab erfolgen hier nicht nur Algorithmenentwicklung und Datensammlung, sondern auch die Modellbildung. Die praktische Berechnung der so definierten Zusammenhänge erfolgt im Anschluss als Datenauswertung und generiert die Ergebnisse zur weiteren Verarbeitung, wie es aus dem Bereich der Statistik bekannt ist. Zu diesen Systemen zählen zum Beispiel der digitale Zulassungsprozess für staatliche Hochschulen, Admission Post Bac (APB) in Frankreich (Lischka und Klingel 2017: 25 ff.) oder die Precobs-Software für ortsbezogene Einbruchsprognosen (a. a. O.: 28 ff.).

Die Unterscheidung nach Lernverfahren zeigt jenseits der technologischen Umsetzung die wesentliche Rolle der Entwickler: Geben sie top-down ein Modell vor oder lassen sie es vom System automatisiert bottom-up entwickeln? Darüber hinaus illustriert sie die Herausforderung beim Entwickeln solcher Systeme: Entwickler beeinflussen die Modellbildung. Direkt erfolgt das beim Top-down-Ansatz, indem sie Optimierungsziele, Erfolgskriterien und Zusammenhänge festlegen. Beim Bottom-up-Ansatz kann die Modellbildung durch die Auswahl der Trainingsdaten indirekt beeinflusst werden.

Abschließend sei betont: Auch bei selbstlernenden algorithmischen Systemen geben Menschen Ziele algorithmischer Entscheidungsfindung, Datenbasis und grundlegende Modelle vor. Bei der Implementierung in einen ADM-Prozess kommen Einsatzbereich, Optimierungsziele und Rahmenbedingungen hinzu. Damit ist die Sorge vor der sogenannten **starken Künstlichen Intelligenz** derzeit nicht gerechtfertigt. Starke Künstliche Intelligenz würde vorliegen, wenn ein selbstlernendes algorithmisches System andere als die vorgegebenen Probleme löst, eigenständig Optimierungsziele definiert und autonom Trainingsdaten auswählt. Starke Künstliche Intelligenz existiert heute nur als Fiktion und Ziel einiger Forschungsvorhaben (Ramge 2018). Wenn einzelne Entscheidungen selbstlernender algorithmischer Systeme schwierig zu erklären sind, resultiert das aus einer Mischung von teilautomatisiertem Lernen und Systemkomplexität – nicht aber aus völliger Automatisierung. Probleme der Nachvollziehbarkeit sollten nicht durch Verweis auf vermeintliche Autonomie sogenannter **Künstlicher Intelligenz** abgetan werden.

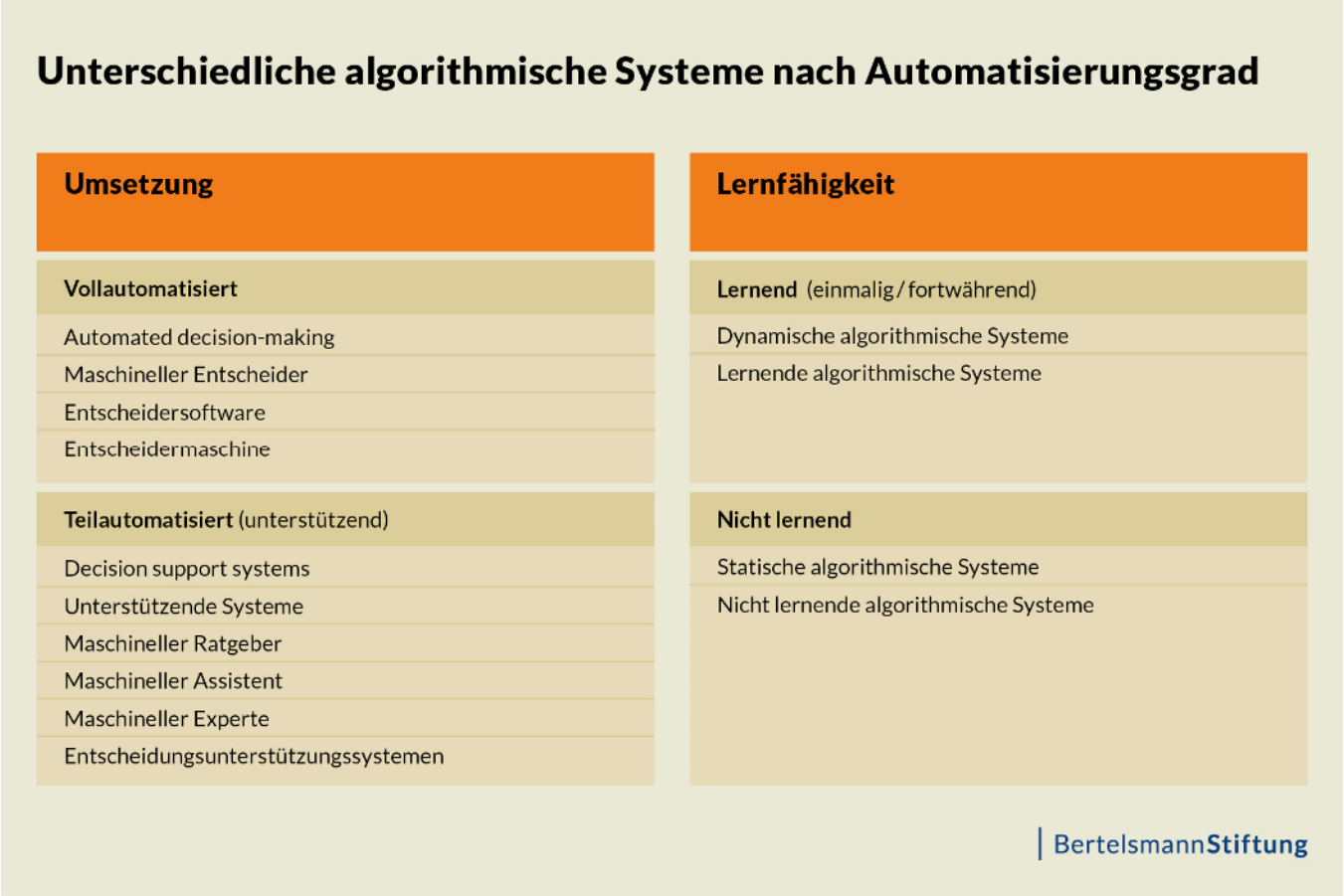
2.1.3 Entscheidungsunterstützung und automatisierte Entscheidungssysteme

Neben den unterschiedlichen Lern- und Entscheidungsprozessen algorithmischer Systeme kann je nach Grad ihrer automatisierten Umsetzung des Outputs in Handlung unterschieden werden in:

- **Assistenzsysteme (Decision Support Systems, DSS-Systeme)**: Sie beraten Menschen, die Entscheidungen abwägen, treffen und umsetzen. Dazu gehört etwa Software, die ortsbezogen die Wahrscheinlichkeit von Einbrüchen vorhersagt. Auf dieser Basis können Polizeistreifen ihre Routen anpassen oder auch nicht.
- **Automatisierte Entscheidungssysteme (Automated Decision-Making Systems, AuDM-Systeme)**: Diese Art von ADM-Systemen setzt auf Basis des Outputs automatisch eine Aktion in Gang. Dazu zählen Bonitätsscoringsysteme, die je nach Bewertung einer Person Zahlungsoptionen (Rechnung, Nachnahme etc.) zulassen oder nicht (Zweig 2018: 11).

Die Implikationen dieser Unterscheidung werden in Kapitel 3.4 weiter ausgeführt.

Abbildung 2:



(Quelle: eigene Darstellung)

2.1.4 Gesellschaftliche Teilhabe

ADM-Prozesse können sich wie menschliche Entscheidungen auf gesellschaftliche Teilhabe auswirken. Sie umfasst im Sinne dieser Expertise die gleichberechtigte Einbeziehung von Individuen und Organisationen in politische Entscheidungs- und Willensbildung sowie die faire Partizipation aller Menschen an sozialer, kultureller und wirtschaftlicher Entwicklung. Es geht also erstens um Teilhabe an demokratischen Prozessen – und damit um politische Gleichberechtigung – und zweitens um Teilhabe an Errungenschaften eines sozialen Gemeinwesens, „angefangen von guten Lebens- und Wohnverhältnissen, Sozial- und Gesundheitsschutz, ausreichenden und allgemein zugänglichen Bildungschancen und der Integration in den Arbeitsmarkt bis hin zu vielfältigen Freizeit- und Selbstverwirklichungsmöglichkeiten“ (Beirat Integration 2013: 1).

Teilhabe in diesem Sinne hat zur Voraussetzung, dass alle Menschen über ein Mindestniveau an materiellen Mitteln verfügen, das ihnen die Mitwirkung am gesellschaftlichen Leben ermöglich. Die Gewährleistung von sozialer und politischer Teilhabe setzt also eine „Sockelgleichheit der sozialen Grundgüter“ (Meyer 2016) voraus. Elemente dieses Sockels werden zum Beispiel in der Allgemeinen Erklärung der Menschenrechte und im Internationalen Pakt über wirtschaftliche, soziale und kulturelle Rechte beschrieben (Bundesgesetzblatt 1966). Um chancengerechte Teilhabe in diesem Sinne zu ermöglichen, sind gezielte Investitionen in die Entwicklung individueller Fähigkeiten nötig (Bertelsmann Stiftung 2011: 31). Es liegt in der Verantwortung des Staates und des Gemeinwesens, jedes Individuum zu befähigen, Chancen tatsächlich nutzen zu können.

Je stärker der mögliche Effekt eines algorithmischen Entscheidungssystems auf Teilhabe ist, desto genauer muss dieses System geprüft werden. Wie sich das Teilhabewirkungspotenzial von algorithmischen Entscheidungssystemen vergleichen lässt, haben Vieth und Wagner (2017) beispielhaft skizziert: Wichtige Anhaltspunkte

ihrer Kategorisierung sind beispielsweise diese Fragen: Werden Menschen bewertet? Wie viel politische und ökonomische Macht hat der Betreiber des algorithmischen Systems? Wie abhängig sind die Bewerteten von dem Ergebnis? Wie groß ist die Reichweite des Systems?

2.2 Gesellschaftliche Anforderungen an algorithmische Prozesse

Wie sehr ADM-Prozesse heute schon den Alltag durchdringen können, sieht man in New York. Die Stadt nutzt sie, um in vielen Lebensbereichen Entscheidungen über ihre Bürger zu treffen. Zum Beispiel, auf welche weiterführende Schule Kinder kommen (Tullis 2014), wo die Polizei wie häufig Streife fährt und kontrolliert (Brennan Center for Justice 2017), ob Lehrer Karriere machen (O'Neil 2017), welche Gebäude vorrangig auf Brandschutz inspiziert werden (Heaton 2015) oder wer des Sozialleistungsbetrugs verdächtigt wird (Singer 2015).

Die Befürworter solcher ADM-Prozesse führen eine Reihe von Chancen an, die sich grob in diese drei Bereiche gliedern lassen (vgl. Lischka und Klingel 2017: 37 f.):

- **Konsistenz**: Algorithmenbasierte Prognosen arbeiten zuverlässig die vorgegebene Entscheidungslogik in jedem Einzelfall ab. Im Gegensatz zu menschlichen Entscheidern ist Software nicht tagesformabhängig und wendet nicht willkürlich in Einzelfällen neue, unter Umständen ungeeignete Kriterien an. Ungeeignete oder gesellschaftlich nicht angemessene Kriterien lassen sich von vornherein ausschließen, die Anwendung im Einzelfall kann detailliert dokumentiert werden. Für die Inkonsistenz und teilweise systematische, diskriminierende Verzerrungen menschlicher Entscheidungen gibt es eine Reihe von empirischen Belegen (vgl. Kahneman et al. 2016), etwa bei der Bewerberauswahl nach ausländisch klingenden Nachnamen (Schneider, Yemane und Weinmann 2014).
- **Komplexitätsbewältigung**: Software kann eine deutlich größere Datenbasis analysieren als Menschen und so Muster erkennen, auf deren Basis bestimmte Aufgaben überhaupt erst oder in einem zuvor nicht möglichen Maß gelöst werden können. Algorithmische Entscheidungssysteme können ihren Output günstig personalisieren und sie lassen sich neuen Umständen leichter anpassen als analoge Strukturen. Der in New York eingesetzte ADM-Prozess zur Schülerverteilung senkte zum Beispiel im ersten Jahr nach Einführung die Anzahl nicht weiterführenden Schulen zugeteilter Schüler von 31.000 auf 3000 (Tullis 2014). Und das System berücksichtigte dabei sowohl die Favoriten der Schüler als auch die Zulassungskriterien der Schulen und die verfügbaren Plätzen, wie der New Yorker Rechnungshof in einer unabhängigen Bewertung zusammenfasst (New York City Independent Budget Office 2016).
- **Effizienz**: Die algorithmische Auswertung großer Datenmengen ist in der Regel günstiger und schneller als die Auswertung vergleichbarer Datenmengen durch Menschen. Eine einmal entwickelte Entscheidungslogik eines Systems lässt sich günstig auf nahezu unbegrenzt viele Fälle anwenden. In New York lobt die Feuerwehr zum Beispiel die Effizienz der zentralisierten algorithmischen Auswertung von Gebäudedaten im Vergleich zum alten papierbasierten und auf 26 Standorte verteilten Verfahren (Heaton 2015). Zudem können algorithmische Entscheidungssysteme in vielen Fällen schneller Output liefern als menschliche Sachbearbeiter.

Die Hoffnung auf höhere Konsistenz, Komplexitätsbewältigung und Effizienz durch algorithmische Entscheidungssysteme bringt Stalder auf den Punkt:

> *„Gerade eine emanzipatorische Politik, die sich angesichts der realen Probleme nicht in die Scheinwelt der reaktionären Vereinfachung zurückziehen will, braucht neue Methoden, die Welt zu sehen und in ihr zu handeln. Und Algorithmen werden ein Teil dieser neuen Methoden sein. Anders lässt sich die stetig weiter steigende Komplexität einer sich integrierenden, auf endlichen Ressourcen aufbauenden Welt nicht bewältigen"* (Stalder 2017: 1).

Der Einsatz von algorithmischen Entscheidungssystemen allein garantiert aber nicht, dass diese Chancen tatsächlich verwirklicht werden. Auch das zeigt die Verwendung solcher Verfahren in New York beispielhaft. Die Risiken für Teilhabechancen durch den Einsatz lassen sich grob in drei Felder unterteilen:

1. Die Optimierungsziele der algorithmischen Systeme
2. Die Umsetzung dieser Ziele in algorithmischen Systemen
3. Die Vielfalt der in einem Bereich genutzten Systeme, Betreiber und Optimierungsziele

Diese drei Felder und übergreifende Rahmenbedingungen beschreiben die folgenden Kapitel.

2.2.1 Gesellschaftliche Angemessenheit der Optimierungsziele

Der Rechnungshof in New York lobt zwar das algorithmische System zur Schülerverteilung, weil es viel mehr Schüler den von ihnen präferierten Schulen zuordnet als das alte Verfahren. Doch zugleich formuliert sein Bericht Zweifel daran, ob die Erfüllung individueller Wünsche wirklich das gesellschaftlich sinnvollste Optimierungsziel eines solchen Zuteilungsverfahrens ist. Das Argument: Unterdurchschnittlich benotete Schüler werden von dem System systematisch unterdurchschnittlich bewerteten Schulen zugeteilt. Das entspricht zwar den Präferenzen der Schüler bzw. ihrer Eltern, benachteiligt aber dennoch Schüler aus ärmeren Vierteln, wo vergleichsweise unterdurchschnittlich bewertete Schulen und unterdurchschnittlich benotete Schüler geballt sind (New York City Independent Budget Office 2016). Hinzu kommt, dass auch Schulen Präferenzen angeben können und einige die Nähe des Wohnorts der Bewerber zur Schule als Hauptkriterium wählen.

Hier geht es nicht um die Effizienz und Konsistenz des algorithmischen Entscheidungssystems, sondern um das Optimierungsziel: Soll das System individuelle Schulpräferenzen in möglichst vielen Fällen befriedigen? Oder soll es die Bildungschancen vom soziodemographischen Hintergrund entkoppeln? Beide Ziele sind vertretbar. Dass ein algorithmisches Entscheidungssystem zuverlässig und nachvollziehbar arbeitet, sagt wenig über seinen gesellschaftlichen Sinn aus. Welches Ziel die Stadt und damit die von ihr beauftragte Technologie verfolgen soll, sollten in einem politischen Willensbildungsprozess möglichst viele Bürger und vor allem potenziell Betroffene mitbestimmen können. Hier geht es um der technischen Umsetzung vorgelagerte Fragen, die nicht anhand von Standardkriterien zu beantworten sind. Die Gestaltung algorithmischer Systeme, die Teilhabechancen berühren, setzt so gut wie immer solche werteorientierten Zieldefinitionen voraus. Was einen guten Arbeitnehmer ausmacht, was eine relevante journalistische Nachricht auszeichnet, woran eine wichtige Freundschaft zu erkennen ist – auf solche Fragen gibt es keine eindeutig richtigen Antworten. Solche sozialen Konzepte müssen die Gestalter von algorithmischen Systemen erst operationalisieren und messbar machen, um diese überhaupt konstruieren zu können. Auch selbstlernende Systeme benötigen von Menschen definierte Optimierungsziele. Dass erfolgreiche Mitarbeiter gesucht werden und wie Erfolg gemessen wird, müssen Menschen entscheiden. Auch wenn ein selbstlernendes System dann später in Datensätzen automatisiert etwa nach korrelierenden Faktoren mit dem Label „erfolgreich“ sucht. Deshalb sind solche Systeme nie autonom:

> *„Es reicht nicht, die Qualität der Werkzeuge zu verbessern, denn Werkzeuge sind nie neutral, sondern reflektieren die Werthaltungen ihrer EntwicklerInnen und AnwenderInnen beziehungsweise deren Auftraggeber oder Forschungsförderer. (…) Was in den technischen Disziplinen unter ‚selbstlernend‘ verstanden wird, ist extrem eng begrenzt: durch Versuch und Irrtum den ‚besten‘ Weg von Punkt A nach Punkt B zu finden, wenn A und B, so wie die Kriterien dafür, was als die beste Lösung anzusehen sei, schon genau definiert sind“* (Stalder 2017: 1).

Allerdings dürfte der gesellschaftliche Diskurs über die Angemessenheit bestimmter Ziele nur in wenigen Fällen allgemeinverbindlich in Gesetzen geronnen kodifiziert sein, wie etwa im allgemeinen Gleichbehandlungsgesetz beispielsweise gegenüber Arbeitgebern. In den meisten Anwendungsfällen algorithmischer Systeme sind verschiedene Wertvorstellungen mit unterschiedlichen Schlussfolgerungen anwendbar. Es gibt nur selten klare, allgemein geteilte Vorstellungen. Daraus folgt: Wenn es keinen Konsens über die gesellschaftliche Angemessenheit der Ziele eines aktuell zu entwickelnden algorithmischen Systems gibt, dann gehört es zur Entwicklung, die

Ziele gesellschaftlich angemessen breit zu diskutieren. Diese Diskussion ist wichtig, um widerstreitende Interessen in jedem neuen Anwendungsfall auszugleichen, zum Beispiel die Interessen von Arbeitgebern, Arbeitnehmern, Arbeitsuchenden. Kein System sollte von einem Interesse bestimmt werden. Damit dieser Ausgleich gelingt, müssen relevante Stakeholder schon in die Entwicklung einbezogen werden.

Die breite Diskussion und das Aushandeln von Optimierungszielen sind notwendig, um gesellschaftlicher Dynamik Raum zu geben. Sonst würde ein automatisch an Daten lernendes algorithmisches System im schlimmsten Fall lediglich die im Trainingsdatensatz gespiegelten gesellschaftlichen Zustände der Vergangenheit fortschreiben. Ein hypothetisches Beispiel: Firma X trainiert ein algorithmisches System zur Bewerberauswahl an den Daten der aktuellen Belegschaft. Optimierungsziel ist es, nur eine bestimmte Anzahl an Kandidaten zu Gesprächen einzuladen, die den erfolgreichsten 20 Prozent der aktuellen Belegschaft am ähnlichsten sind. Ein solches System wird höchstwahrscheinlich die soziodemographische Zusammensetzung der erfolgreichsten 20 Prozent der aktuellen Belegschaft reproduzieren – und damit die bislang durch systematische Verzerrungen in menschlicher Entscheidungsfindung geschaffene Verteilung. Denn menschliche Entscheidungen sind nicht per se fairer als algorithmische und in einigen Einsatzbereichen nachweislich unfair. Sie sind durch Vorurteile und teilweise ungeeignete Kriterien bestimmt. Dafür gibt es eine Reihe von empirischen Hinweisen. Bei der Bewerberauswahl haben zum Beispiel fremd klingende Namen einen solchen verzerrenden Effekt:

> *„Um eine Einladung zum Vorstellungsgespräch zu erhalten, muss ein Kandidat mit einem deutschen Namen durchschnittlich fünf Bewerbungen schreiben, ein Mitbewerber mit einem türkischen Namen hingegen sieben“* (Schneider, Yemane und Weinmann 2014: 4).

Bei dem algorithmischen System zur Schülerverteilung in New York wirkt vermutlich ein ähnlicher Effekt der Status-quo-Reproduktion: Die Standorte überdurchschnittlich erfolgreicher Schulen sind aufgrund der Verteilung von Reichtum und Bildungsniveau im Stadtgebiet häufig in räumlicher Nähe zu den Wohnorten von Schülern aus reicheren Haushalten. Die räumliche Nähe beeinflusst wiederum die Präferenzen der Schüler und das Verteilungsergebnis, unabhängig vom algorithmischen System zur Auswahl.

2.2.2 Umsetzung der Ziele in Systemen

Gut gemeint ist nicht gut gemacht: Auch algorithmische Entscheidungssysteme mit gesellschaftlich angemessenen Optimierungszielen können teilhabemindernde Effekte haben, weil es bei der Umsetzung hapert. Um die Qualität eines algorithmischen Systems zu beurteilen, muss es im Einsatz untersucht werden.

Die Umsetzungsqualität ist auf vielerlei Weise mit der Zieldefinition verknüpft. Anschaulich wird dies am Beispiel des in einer Pilotstudie am Berliner Bahnhof Südkreuz getesteten Überwachungssystems. Ein algorithmisches System gleicht dort die Überwachungsvideos mit Fotos gesuchter Personen ab, um polizeilich Gesuchte zu erkennen. Bei der Umsetzung eines solchen Systems müssen mindestens zwei Optimierungsziele abgewogen werden: Einerseits möglichst viele Gesuchte korrekt in der Menge aller Aufgenommenen identifizieren (hohe Sensitivität). Oder andererseits möglichst wenige Unschuldige in der Menge aller Aufgenommenen fälschlicherweise als Gesuchte erkennen (hohe Spezifizität). Beide Ziele können nicht gleichzeitig maximiert werden. Höhere Sensitivität geht mit niedrigerer Spezifizität einher und umgekehrt. Oder am Beispiel erklärt: Wenn das Erkennungssystem nahezu alle Verdächtigen korrekt erkennen soll, müssen auch täglich viele unschuldige Bürger festgehalten und erkennungsdienstlich behandelt werden. Bei 160.000 Passanten am Tag und einer Quote von einem Prozent fälschlich als „gesucht“ Erkannten würde es rund 1600 ungerechtfertigte Fehlalarme und Personenkontrollen am Tag geben. Man kann ein System darauf (hohe Sensitivität) optimieren. Ob das eine Gesellschaft will, ist eine andere Frage. Hier ist die Umsetzung eng mit Zieldefinition und -priorisierung und den geltenden rechtlichen Beschränkungen verbunden: „Die geeignete Parametrierung eines solchen Systems erfordert eine Güterabwägung und kann eine politische Fragestellung sein“ (Gallwitz 2017:1)

Ein bekanntes Beispiel für die Analyse der Umsetzungsqualität eines algorithmischen Entscheidungssystems ist die 2016 veröffentlichte Studie der US-Rechercheorganisation Propublica zur Qualität algorithmischer Rückfallprognosen, die in vielen US-Bundesstaaten vor Gericht genutzt werden. Die Software war zu diesem Zeitpunkt seit Jahren im Einsatz, doch zuvor hatte niemand systematisch überprüft und öffentlich gemacht, welche Fehler

bei den Prognosen auftreten. Kernergebnis der Propublica-Studie: Die Art der Fehlprognosen unterscheidet sich zwischen schwarzen und weißen Personen. Der Anteil Schwarzer mit hoher Rückfallprognose, aber ohne Rückfall binnen zwei Jahren ist doppelt so hoch wie bei Weißen (Angwin et al. 2016: 2). Erst diese Rechercheergebnisse brachten eine Diskussion über Fairnesskriterien der Systeme in Gang. Es braucht also systematische Untersuchungen der tatsächlichen Entscheidungsqualität teilhaberelevanter Systeme.

In New York hat die fehlende Nachvollziehbarkeit eines algorithmischen Systems zur Bewertung von Lehrern dazu geführt, dass ein Gericht den Einsatz dieser Software untersagte. Das System habe „willkürliche“ und „unbeständige“ („arbitrary and capricious“) Ergebnisse geliefert, hieß es in der Urteilsbegründung (Harris 2016). Fehlende Überprüfbarkeit und Nachvollziehbarkeit sind auch die Hauptkritikpunkte an einem algorithmischen System zur Planung von Streifenfahrten der New Yorker Polizei. Der Stadtrat James Vacca formuliert seine Bedenken so: Die Polizei habe ihm als Volksvertreter die Kriterien und Entscheidungslogik für die Einsatzplanung in der Bronx nie hinreichend erklären können: „That always annoyed me, and I felt that I was not being given a lot of the answers I wanted” (Powles 2017: 1).

2.2.3 Vielfalt der Systeme und Betreibermodelle

Die gesellschaftliche Angemessenheit der Optimierungsziele und die Qualität der Umsetzung sind immer an einzelnen algorithmischen Systemen zu bewerten. Doch es gibt auch auf der darüber liegenden Ebene, der Gesamtheit aller Systeme, Handlungsbedarf. Eine große Vielfalt an Systemen und Betreibermodellen ist ein Wert an sich im Hinblick auf Teilhabe. Vielfalt nach unserem Verständnis umfasst:

- **Vielfalt der Ziele und Betreiber**: Unterschiedliche Optimierungsziele in einem Einsatzgebiet. Damit verbunden: Vielfalt von Auftraggebern und Betreibern von algorithmischen Entscheidungssystemen – also zum Beispiel Betreibern aus dem öffentlichen, dem privatwirtschaftlichen wie auch dem zivilgesellschaftlichen Sektor mit ihren unterschiedlichen Ansätzen und Organisationszielen.
- **Vielfalt der Umsetzung und der Systeme**: Unterschiedliche Operationalisierungen in einem Einsatzgebiet.

Die Vielfalt von algorithmischen Entscheidungsprozessen ist herausgefordert, denn algorithmische Entscheidungssysteme schalten insbesondere im Kontext von Technologien Künstlicher Intelligenz „der Monopolisierung den Turbo zu“ (Ramge 2018: 88). Dabei wirken zwei Tendenzen zusammen: Zu beachten sind einerseits die Netzwerkeffekte digitaler Plattformen, Infrastrukturen sowie Hard- und Softwaresysteme, in deren Folge Unternehmen wie Microsoft, Apple, Amazon, Google und Facebook sowie Yandex, Tencent, Baidu und Alibaba Oligopolstrukturen im Bereich Datenwirtschaft entwickelten. Andererseits basiert die Entwicklung von selbstlernenden algorithmischen Entscheidungssystemen auf dem Vorhandensein und der Nutzung von Feedbackdaten, die ebenfalls vor allem den Unternehmen zur Verfügung stehen, die bereits im Markt aktiv sind. „Je öfter sie [die Feedbackdaten, Anm. d. Verf.] genutzt werden, je mehr Marktanteile sie erobern, desto schwerer wird ihr Vorsprung aufzuholen sein“ (ebd.).

Für aktive Vielfaltssicherung bei algorithmischen Entscheidungssystemen spricht:

- **Konzentrationstendenz durch Skalierbarkeit**: Die einmal entwickelte Entscheidungslogik eines algorithmischen Systems ist auf sehr viele Fälle anwendbar, ohne dass die Kosten für den Einsatz substanziell steigen. Das führt dazu, dass in einigen Lebensbereichen wenige algorithmische Systeme dominieren können. Die Tendenz zur Konzentration der Macht ist bei ADM-Systemen größer bei anderen Strukturen. Je geringer die Vielfalt algorithmischer Systeme in einem Einsatzbereich ist, desto härter treffen Fehler in der Umsetzung die Betroffenen. Je größer die Reichweite eines algorithmischen Systems ist, desto schwieriger ist es für den Einzelnen, sich den Verfahren und Folgen zu entziehen.
- **Abbilden gesellschaftlicher Pluralität und Dynamik**: Für soziale Phänomene oder Konzepte, wie zum Beispiel Nachrichtenrelevanz oder Eignung von Bewerbern, existieren viele kontextabhängige Operationalisierungen. Solche Konzepte unterliegen dem gesellschaftlichen Wandel. Je geringer die Vielfalt

algorithmischer Systeme in einem Einsatzbereich ist, desto kleiner wird der Raum für die Abbildung gesellschaftlicher Pluralität und Dynamik – zum Beispiel durch unterschiedliche Optimierungsziele.

- **Raum für Innovation:** Wenn in einem Einsatzfeld unterschiedliche algorithmische Systeme im Einsatz sind, kann der Vergleich zwischen ihnen Erkenntnisse über Wirkung, Fehlerquellen und Alternativen befördern. Das ist die Grundlage für Innovation bei der Umsetzung und damit gesellschaftlichen Fortschritt.

2.2.4 Schaffung übergreifender Rahmenbedingungen für teilhabeförderliche Systeme

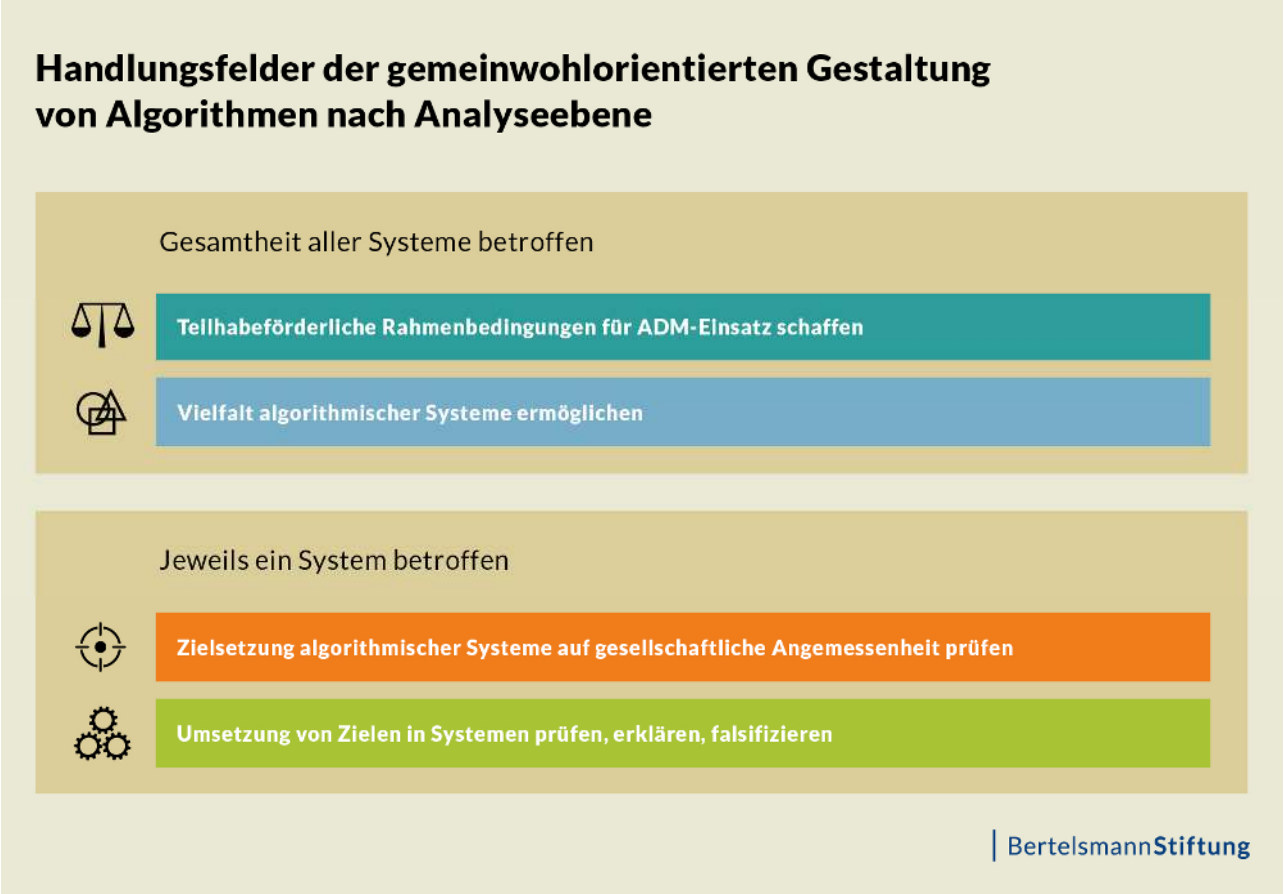
Die Angemessenheit der Optimierungsziele, die Qualität der Umsetzung und die Vielfalt der algorithmischen Systeme und Betreibermodelle – der Handlungsbedarf auf diesen drei Feldern führt zu einem vierten: Es braucht kompetente Akteure, um den Rahmen für eine positive Entwicklung zu gestalten. Individuelle und staatliche Kompetenz verstehen wir als eine wesentliche übergreifende Rahmenbedingung. Denn wie es der Kosmologe und Mitgründer des Future of Life Institute, Max Tegmark, formuliert: Gesellschaftlicher Nutzen stellt sich nicht von alleine ein:

> *„I'm optimistic that we can create an inspiring future with AI if we win the race between the growing power of AI and the growing wisdom with which we manage it, but that's going to require planning and work, and won't happen automatically“* (Torres 2017: 1).

Positive, gemeinwohlfördernde Gestaltung umfasst einerseits staatliche Abwehr und Nachsorge durch Regulierung. Anderseits ist der Staat aber auch als aktiver Gestalter und Ermöglicher gefragt. Es ist Aufgabe des öffentlichen Sektors, den gesellschaftlich sinnvollen Einsatz algorithmischer Systeme in der Daseinsvorsorge zu fördern – ohne dass Partikularinteressen etwa von Investoren oder Dienstleistern aus der Wirtschaft dominieren. Handlungsbedarf besteht hier offenkundig bei der staatlichen Gestaltungskompetenz. Das gilt nicht nur für das Beispiel New York, wo das Stadtparlament Ende 2017 beschloss, einen Arbeitsstab einzurichten, um die Qualität der von der Stadt genutzten algorithmischen Systeme zu untersuchen. Die Juristin Julia Powles zeigt in einer ersten Bewertung des Vorhabens, dass dieser Arbeitsstab Kompetenzen in beiden Wortbedeutungen braucht, um seine Aufgabe erfüllen zu können: Sachverstand und Fähigkeiten einerseits, rechtliche Handhabe und Zuständigkeit andererseits:

> *„There is no readily accessible public information on how much the city spends on algorithmic services, for instance, or how much of New Yorkers' data it shares with outside contractors. Given the Council's own struggle to find answers, the question now is whether the task force will do any better. Can it develop good recommendations, and fulfill its mandate, without the close cooperation of agencies and contractors? (…) The law's second apparent failing is that it doesn't address how the city government, and those who advise it, can exercise some muscle in their dealings with the companies that create automated-decision systems“* (Powles 2017: 1).

Abbildung 3:



(Quelle: eigene Darstellung)

3 Was zu berücksichtigen ist: Herausforderungen algorithmischer Systeme

Das vergangene Kapitel beleuchtete die begrifflichen Grundlagen von algorithmischen Entscheidungsprozessen und gab einen Überblick über Anwendungsbereiche und gesellschaftliche Anforderungen.

Das folgende Kapitel vertieft das Verständnis der Komplexität bei der Gestaltung, Anwendung und Bewertung von algorithmischen Entscheidungsprozessen und zeigt besondere Herausforderungen auf bei der Implementierung von ADM-Systemen in den gesellschaftlichen Kontext. Dieses Arbeitspapier fokussiert auf Prozesse und Strukturen, die gesellschaftliche Teilhabe beeinflussen (vgl. zu Teilhaberelevanz Vieth und Wagner 2017). Auf andere ADM-Prozesse wird hier deshalb nur zum Zwecke der Abgrenzung oder Illustrierung verwiesen. Je größer die Teilhaberelevanz, desto größer auch die Anforderungen an die Sicherung gesellschaftlicher Angemessenheit, Überprüfbarkeit der Funktionsweise sowie Diversität

3.1 Einsatzfeld: Sind teilhaberelevante Fragen berührt?

Dieses Arbeitspapier bezieht sich auf ADM-Prozesse, die die chancengerechte Teilhabe im hier verstandenen Sinne betreffen. Bereits diese Einordnung birgt Herausforderungen:

Das auf gleiche **„Verwirklichungschancen“** ausgerichtete Konzept gesellschaftlicher Teilhabe fokussiert auf die staatliche Verantwortung in der Gewährleistung von Teilhabechancen für alle Individuen, unabhängig von sozialem Hintergrund oder der Zugehörigkeit zu bestimmten gesellschaftlichen Gruppen. Individuen sollen kontinuierlich dazu befähigt werden, ihre individuellen Chancen zu nutzen. Dabei kommt dem gleichen Zugang zu Bildung und Beschäftigung eine übergeordnete Rolle zu (Bertelsmann Stiftung 2011), die von gleichem Zugang zu sozialer Sicherung, Gesundheitsversorgung und Freizeitgestaltung begleitet wird (Beirat Integration 2013).

Algorithmische Prozesse versprechen einerseits, durch ihre Anwendung auf möglichst große Datenbestände (Big Data) für den jeweiligen Anwendungsbereich möglichst „passgenaue“, d. h. im besten Fall personalisierte Resultate zu liefern. Dies kann Teilhabechancen verbessern, etwa im Bereich von Bildung. Hier verspricht der Einsatz intelligenter Softwarelösungen, den individuellen Lernfortschritt zu unterstützen (Dräger und Müller-Eiselt 2015; Stone et al. 2016). Auch die Einführung von algorithmischen Entscheidungssystemen bei Arbeitsagenturen in Polen orientiert sich an der Verbesserung individueller Serviceleistungen für Erwerbslose (Jedrzej, Sztandar-Sztanderska und Szymielewicz 2015: 8).

Andererseits kann der Zugang zu Bildungs- oder Beschäftigungschancen unter Umständen auch versperrt werden, wenn Algorithmen zum Einsatz kommen. Die effektive Unterscheidung ist eines ihrer wesentlichen Merkmale. Wie Barocas und Selbst (2014) am Beispiel der automatisierten Bewerberauswahl illustrieren, besteht der Clou der ADM-Systeme darin, eine rationale Basis für die Klassifizierung von Bewerbern herzustellen – auf Grundlage von Kriterien, die sich in der Vergangenheit bewährt haben. Dabei kann es neben der **intendierten Selektion** auch leicht zu **nicht intendierter Diskriminierung** kommen, wie sich bereits bei der Definition des Ziels und seiner Operationalisierung zeigt. Stellte beispielsweise die Dauer eines Beschäftigungsverhältnisses ein zentrales Kriterium für die Bewerberauswahl dar, wäre eine algorithmische Entscheidung gegen die Einstellung von Frauen folgerichtig – denn sie verlassen aufgrund der Geburt eines Kindes statistisch gesehen häufiger eine Arbeitsstelle.

Jenseits der Zielstellung und Operationalisierung stellen **Trainings- und Analysedaten** des ADM-Systems eine vielfältige Quelle für Diskriminierung dar. Oft spiegeln sie gesellschaftlich institutionalisierte Ungleichheiten wider. Regelmäßig sind marginalisierte Gruppen über- oder unterrepräsentiert (a. a. O.). Der Rückgriff auf diskriminierendes Datenmaterial zur Vorhersage von Entwicklungen birgt die Gefahr der Reproduktion von Diskriminierung.

Eine solche lässt sich kaum durch die Entfernung einzelner Variablen umgehen, die auf eine bestimmte Gruppenzugehörigkeit verweisen – sogenannte **sensitive Attribute**. Denn mit großer Wahrscheinlichkeit spiegelt sich das Attribut auch in anderen Daten, die damit in Verbindung stehen. Hierbei handelt es sich um sogenannte **Proxy-Variablen**, wie beispielsweise die Postleitzahl des Wohnorts. Denn Wohnort und sozioökonomischer Status korrelieren häufig ebenso wie sozioökonomischer Status und ethnische Herkunft.

Dazu kommt: Die Frage der Diskriminierung betrifft im Regelfall nicht nur die **Repräsentativität der Daten**, sondern gleichfalls ihre **Klassifikation und Variablenauswahl**:

> *„The next biases come from the training data. A data mining system learns by example, and must take its training data as ‚ground truth', as that data is the only information the algorithm has about the world outside. A big part of getting the data right is correctly labeling the examples that the algorithm is trained on. The most common source of data for predictive policing algorithms – used in every version of predictive policing in existence – is past crime data, often collected by the police themselves. (…) Reliance on past data is a big problem, though, as accurate crime data often does not exist. There are several reasons for this, but one major one is that the most systematic contact police departments have with ‚criminals' is at the moment of arrest. Results after arrest are often not updated. Thus, most research in crime statistics uses arrest data as the best available proxy, even though arrests are racially biased. (…) As a result, a good number of the crime labels may be incorrect, (…) Training data must also be a representative sample of the whole population. The ultimate goal of data mining is pattern-matching and generalization, and without a representative sample, generalizing introduces sampling bias. There are many potential sources of sampling bias. The data can be skewed by past historical practices, for example (…) Another source of discriminatory effect is feature selection"* (Selbst 2016: 17–20).[4]

Die Gewährleistung eines Datensets, das zur Erreichung bestimmter Ziele eines algorithmischen Entscheidungssystems geeignet ist, stellt eine Herausforderung dar. Ein Beispiel: Will der Entwickler eines ADM-Systems die Auswahl von Bewerbern nach Hautfarbe ausschließen, müssen auch Merkmale in den Trainingsdaten erkannt werden, die mit der Hautfarbe korrelieren (z. B. Wohnort).[5] Nur so kann eine Diskriminierung über Proxy-Werte erkannt und verhindert werden. Dem Umgang mit möglicher Diskriminierung bzw. der Herstellung von Teilhabegerechtigkeit kommt eine wesentliche Bedeutung im gesamten Entwicklungs- und Anwendungsprozess vorhersagebasierter Analysesysteme zu. Letztlich stellt sich die Frage: Wann ist Diskriminierung teilhaberelevant? Wann bedarf es einer bestimmten Vielfalt oder Diversität, die in das **Design von algorithmischen Entscheidungssystemen** integriert ist?

Vieth und Wagner (2017) haben auf diese Frage eine instruktive Antwort gegeben. Sie orientierten dabei die **Teilhaberelevanz** nicht an dem **Anwendungsbereich**, sondern an dem **Einfluss und der Reichweite des ADM-Prozesses**. Entscheidende Variablen sind dabei a) die politische und ökonomische Macht bzw. Marktstellung eines Betreibers, b) die Existenz alternativer Entscheidungsverfahren und/oder Produktangebote sowie c) das Verhältnis von algorithmischer Determination und menschlicher Entscheidungs- und Handlungsautonomie. In der

---

[4] Das Zitat illustriert einige Probleme und Lösungsansätze im Bereich adäquater Datenauswahl und -analyse. Eine vollständige Auflistung würde den Rahmen dieses Arbeitspapieres sprengen. Allerdings ist anzumerken, dass falsche Kategorisierungen von Daten bei der späteren Verarbeitung Berücksichtigung finden können. Zudem ist die Vermutung über die Wahrscheinlichkeitsverteilungen der Daten und Zielgrößen zu berücksichtigen, die Eingang in die Auswahl von Analyseverfahren bzw. Algorithmen finden (vgl. Bayes'sche Verfahren, Russel und Norvig 2012).

[5] Unter Umständen können selbst solche Variablen zur Vorhersage gruppenbezogener, sensibler Merkmale geeignet sein, die für sich genommen keine Korrelation mit der Zielgröße haben. Um wirklich ausschließen zu können, dass ein Verfahren diese Informationen extrahieren kann, muss unter Nutzung des gleichen Verfahrens ein Training durchgeführt werden. Hier wird mit den ursprünglichen Eingabedaten versucht, das sensitive Attribut vorherzusagen. Nur wenn dies nicht gelingt, können die Daten als diskriminierungsfrei eingeschätzt werden. Zusätzlich müssen auch die Ergebnisse des Algorithmus überwacht werden, wozu das sensitive Attribut (also z. B. die Hautfarbe) wiederum gebraucht wird (ST-IFT).

Praxis bedeutete dies, dass der Frage von Diversität insbesondere dann eine hohe Bedeutung zukommt, wenn der ADM-Prozess einen großen und kaum zu umgehenden Einfluss auf menschliches Handeln ausübt.

Dennoch wird die Frage, welches **Maß an Diversität in der Gesellschaft** wünschenswert ist bzw. wie man Teilhabegerechtigkeit fördert, eine andauernde **normative Herausforderung** darstellen – die in den verschiedenen Anwendungsbereichen wie Gesundheit, Bildung oder Sicherheit spezifische Antworten benötigt.

3.2 Zielsetzung und Evaluation: Wer definiert und kontrolliert Erfolg – und wie?

Wie bereits betont, kommt der Zielstellung algorithmischer Systeme mit Teilhaberelevanz eine übergeordnete Bedeutung zu. Sie soll gemeinwohlförderlichen Zielen wie der Verbesserung der Gesundheitsversorgung oder der schulischen Bildung dienen. Damit muss nachvollziehbar sein, ob algorithmische Analyse- und Entscheidungssysteme diesen Anforderungen gerecht werden. Einerseits ist davon auszugehen, dass selbstlernende Algorithmen bzw. Technologien Künstlicher Intelligenz in Bereichen wie der Gesundheitsversorgung, öffentlichen Sicherheit oder sozialen Sicherung neuartige Erkenntnisse sowie effiziente Entscheidungs- und maßgeschneiderte Handlungsoptionen ermöglichen (Stone et al. 2016). Andererseits bergen sie die bereits aufgezeigten wie auch weitere Risiken (vgl. Kapitel 3.5). Dazu zählen der Zielstellung nicht entsprechende Systemdesigns (bspw. fehlende Vergleichbarkeit von Trainings- und Realdaten bei wesentlichen Eigenschaften, fehlendes Feedback), unangemessene Einsatzszenarien oder schlicht und einfach: mehr oder weniger komplexe Anwendungsfehler (Diakopoulos 2016; Future of Privacy Forum 2017; Kroll et al. 2017).

In der wissenschaftlichen Debatte besteht Konsens darüber, dass Künstliche Intelligenz dem Menschen bzw. der Gesellschaft dienen soll. Dennoch gibt es **bislang nur allgemeine unverbindliche Leitlinien** – etwa zu Sicherheit, Transparenz, dem Erhalt menschlicher Handlungsautonomie oder sozialer Gerechtigkeit (Calo 2017; Cave 2017; FAT/ML 2016; Future of Life Institute 2017; Georgieva 2017). Die Debatte um Zielsetzung und Evaluation steht noch am Anfang. Die ersten Leitlinien wurden bislang nicht zur verbindlichen Bewertung strittiger Fälle genutzt, daher fehlt die **Konkretisierung** ebenso wie die **Verortung von spezifischen und transparent geregelten Verantwortlichkeiten**. Nimmt man algorithmische Entscheidungssysteme in ihrer Gesamtheit in den Blick, stellen sich hinsichtlich Zielsetzung und Evaluation mindestens folgende Fragen:

- Wer definiert die Ziele automatisierter Entscheidungsverfahren? Entsprechen sie demokratischen Grundsätzen? Können Stakeholder einbezogen werden?
- Wann müssen die kodifizierten Ziele automatisierter Entscheidungsverfahren transparent gemacht werden und ggf. wem gegenüber?
- Wer verantwortet die Umsetzung automatisierter Entscheidungsverfahren? Entsprechen Design und Implementierung den Zielen des automatisierten Entscheidungsverfahrens bzw. unter welchen Bedingungen kann von dieser Kongruenz abgewichen werden?
- Wer bestimmt die Art und Häufigkeit der Evaluation? Findet sie Eingang in die Systemweiterentwicklung?

Die Antworten auf diese Fragen dürften recht komplex ausfallen. Ein Beispiel zur Veranschaulichung aus dem Gesundheitsbereich: Algorithmische Systeme könnten zwecks personalisierter Diagnose und Behandlung entwickeln werden. Schon die sehr abstrakten Ziele sind schwierig abzuwägen: Therapieoptimierung? Effizienteste Auslastung von Krankhauskapazitäten? Kostensenkung? Qualitätsmaximierung in jedem Einzelfall? (Executive Office of the President et al. 2014; Prainsack 2017; Stone et al. 2016). Beim Gestalten und Implementieren algorithmischer Entscheidungssysteme treffen unterschiedliche Stakeholder wesentliche Wertentscheidungen.

3.3 Dynamik und Komplexität: Wie entwickelt sich das Entscheidungssystem?

Die aus der US-Debatte bekannten ADM-Systeme in Bereichen wie Bewerberauswahl, Kriminalitätsprognosen oder Delinquenzrisiken sind umstritten. Zentrale Kritikpunkte betreffen ihre

- **Unsichtbarkeit** (Betroffenen ist ihre Existenz zumeist nicht bekannt),
- **fehlende Transparenz** (die Funktionsweise, Algorithmen oder die ihnen zugrunde liegende Logik unterliegen häufig Geschäftsgeheimnissen),
- **hohe Reichweite** (ein Fehler betrifft unter Umständen eine hohe Zahl an Menschen) und
- **Zweckentfremdung** (Systeme werden zu anderen Zwecken eingesetzt als von ihren Entwicklern vorgesehen).

Als solche wurden sie als „Weapons of Math Destruction“ (O’Neil 2016) – statistische Massenvernichtungswaffen – charakterisiert. O’Neill pointierte, dass die scheinbar neutralen ADM-Systeme fehlerhaft entwickelt, implementiert oder angewandt sein können. Aufgrund der Probleme von Unsichtbarkeit, fehlender Transparenz und der hohen Reichweite könne dies allerdings unter Umständen lange verborgen bleiben. Darunter leiden Betroffene – häufig ohnehin aus marginalisierten Bevölkerungsgruppen – erheblich. Allerdings: Würden die Betreiber der ADM-Systeme eine qualifizierte Transparenz über die Systeme ermöglichen, d. h. über zugrunde liegende Daten, Modelle und Algorithmen, wäre eine Evaluation der Systeme möglich (Zweig 2016).

Die in vielen US-Staaten vorgenommene Berechnung von Kriminalitätsprognosen vor Gericht durch das sogenannte **COMPAS-System (Correctional Offender Management Profiling for Alternative Sanctions)** basiert beispielsweise auf Fragebögen, deren Auswertung Grundlage für individuelle Scoringwerte ist. Die Anzahl der automatisierten Entscheidungen zugrunde liegenden Variablen ist überschaubar. Kritik ergibt sich aus der fehlenden Transparenz über die Gewichtung der Variablen (Funktionsweise) und der unzureichenden Evaluation (z. B. hinsichtlich des Zusammenhangs zwischen Delinquenzrisiko und tatsächlicher Rückfallquote). Besonders hohe Delinquenzprognosen sind für farbige Menschen dokumentiert (Angwin et al. 2016).

Ähnlich verhält es sich mit den bekannten personen- oder ortsbezogenen Kriminalitätsprognosen, die auf Basis vorhandener Daten über vorbestrafte Bürger bzw. geographische Kriminalitätsschwerpunkte erstellt werden und die der Prävention oder Ermittlung dienen. Sie sind in ihrer Wirksamkeit, korrekten Anwendung und sozialen Wechselwirkung umstritten, was durch mangelnde Transparenz bestärkt wird (Lischka und Klingel 2017). Doch auch diese dynamischeren Verfahren bleiben trotz aller Kritik in ihrer Funktionsweise und Wirkung überschaubar.

Anders dagegen sind Systeme zu bewerten, die auf eine unbekannte und/oder unüberschaubare Menge an Daten zurückgreifen, die Datenanalysetechnologien aus dem Bereich der Künstlichen Intelligenz bzw. des maschinellen Lernens anwenden und einer enormen Dynamik ausgesetzt sind.

Hier können die unterschiedlichsten Daten und Datenanalysesysteme, wie Text-, Bild- und Aktivitätsanalysen bzw. Mustererkennung, miteinander verknüpft und integriert werden. Solche **komplexen, auf hoch dynamischen Interaktionen basierenden algorithmischen Entscheidungssysteme** kommen beispielsweise bei personenbezogenen Kriminalitätsprognosen auf Basis der Aktivitäten in sozialen Netzwerken zum Einsatz (im Gegensatz zu einfachen personen- oder ortsbezogenen Prognosen, siehe Selbst 2016). Trotz fehlender Übersicht und Evaluation (Ferguson 2017) finden sie Medienberichten zufolge international Verbreitung (Dahllof et al. 2017; Dickey 2016; Mateescu et al. 2015). Sie stellen die Prüfung gesellschaftlicher Angemessenheit und die Evaluation ihrer Wirkung in dreifacher Hinsicht vor Herausforderungen: Sie basieren auf selbstlernenden Algorithmen, sind eingebunden in ein hoch dynamisches System und unterliegen der Kontrolle der sie hervorbringenden Plattformen in privater Hand.

Das Potenzial dieser hoch dynamischen algorithmischen Analyse- und Entscheidungssysteme lässt sich mit Blick auf aktuell existierende Datenbestände (Christl 2014; 2017) und Analysekapazitäten nur erahnen.[6] Sicher ist allerdings: Mit der Komplexität der einem Entscheidungssystemen zugrunde liegenden Datenbestände und

[6] Instruktive Beispiele finden sich insbesondere im Bereich der Privatwirtschaft, beispielsweise in der Aufdeckung statistischer Zusammenhänge zwischen dem Umgang mit Auto-Complete-Funktionen in privaten Messangerdiensten und der Lebenserwartung.

Analyseverfahren steigt auch dessen Dynamik und Komplexität. Nachvollziehbarkeit und Kontrolle werden schwieriger. Deshalb verschiebt sich der Fokus von der Prüfung des Algorithmus auf die Überprüfung des Gesamtsystems und der daraus hervorgehenden Entscheidungen, beispielsweise die Angemessenheit der Datenbasis, Modellierung und Input-Output-Beziehungen (vgl. Kapitel 4.2).

Je nach Anwendungsbereich von ADM-Prozessen ist anzumerken, dass eine fehlende Nachvollziehbarkeit und Kontrolle verfassungsrechtliche Bedenken hervorruft: Gemäß unserer freiheitlich-demokratischen Grundordnung liegt die Gesetzgebungskompetenz bei der Legislative. Gesetze werden durch eine mittelbar demokratisch legitimierte Exekutive ausgeführt. Beide Gewalten unterliegen der Kontrolle durch die Wählerinnen und Wähler. Diese Legitimationskette bricht ab, wenn der Algorithmus zur „Black Box“ wird, der sich jeder Kontrolle entzieht.

3.4 Automatisierung: Wie eigenständig agiert das Entscheidungssystem?

Die Einführung automatisierter Entscheidungen stellt das Recht vor Herausforderungen: Die ab Mai 2018 zur Anwendung kommende Europäische Datenschutz-Grundverordnung (DSGVO-EU) schreibt etwa vor, dass eine Person generell das Recht hat, „nicht einer ausschließlich auf einer automatisierten Verarbeitung (…) beruhenden Entscheidung unterworfen zu sein, die ihr gegenüber rechtliche Wirkung entfaltet oder sie in ähnlicher Weise beeinträchtigt“ (Europäisches Parlament und Rat der Europäischen Union 2016). Weiterhin sieht sie bestimmte Informations- und Transparenzpflichten für die Betreiber automatisierter Entscheidungssysteme vor, die Auskunft zur Logik des Entscheidungssystems, seiner Tragweite und der angestrebten Auswirkungen beinhalten (a. a. O.).[7] Die Norm klingt zunächst eindeutig. Doch wann ist eine Entscheidung automatisiert, wann gelten diese Regeln?

Es lassen sich grob zwei Typen von algorithmischen Systemen unterscheiden. Einige Systeme bereiten Entscheidungen vor. Sie schaffen die Basis, auf der dann Menschen entscheiden. Zum Beispiel Richter in US-Bundesstaaten, denen eine per Software erstellte Prognose des Delinquenzrisikos der Angeklagten vorliegt. Solche algorithmischen Systeme, die Entscheidungen vorbereiten, etwa die COMPAS-Delinquenzprognosen oder die Precobs-Software für ortsbezogene Einbruchsprognosen (Lischka und Klingel 2017: 28 ff.) verstehen wir als **Assistenzsysteme** oder **Decision Support Systems (DSS)**.

Andere algorithmische Systeme setzen Entscheidungen automatisch um. Im Folgenden wird eine Software als **automatisches Entscheidungssystem** oder **Automated Decision-Making System (AuDM-System)** bezeichnet, wenn ein algorithmisches System eine Bewertung oder Prognose ausgibt und diese von einer Software unmittelbar in eine Entscheidung umgesetzt wird. Beispielsweise, wenn Software nach einer Fallprüfung bei Verdachtsmomenten automatisch Mahnungen verschickt wie beim australischen Centrelink-System (Rohde 2017).

In der Verwaltungspraxis sind die unterschiedlichsten Modelle bekannt, sowohl hinsichtlich der Entscheidungsautonomie als auch daran gekoppelter Widerspruchsrechte (Citron 2008: 1263 ff.): Hoch automatisierte Systeme ohne menschliche Entscheider sind beispielsweise die Bewerbervorauswahl auf Basis von Onlinepersönlichkeitstests im angloamerikanischen Raum oder die Studienplatzvergabe in Frankreich (Lischka und Klingel 2017). Demgegenüber sind die in dem US-amerikanischen COMPAS-System automatisiert hergestellten Prognosen zum Delinquenzrisiko von Straftätern nur eine von verschiedenen Variablen, auf deren Basis ein Richter entscheidet (a. a. O.). Hier handelt es sich um ein Assistenzsystem, bei dem ein bestimmtes Maß an menschlicher Handlungsautonomie integriert ist. Gleiches gilt für die Software, die polnische Arbeitsämter bei der Jobvermittlung unterstützt. Sie unterteilt Erwerbslose in verschiedene Kategorien, die mit verschiedenen Unterstützungsprogrammen korrespondieren (Vieth und Wagner 2017).

[7] Der Bezug zur Datenschutz-Grundverordnung wurde hergestellt, da die Norm aktuell von größter Bedeutung ist. Eine umfassende Rechtsanalyse und -auslegung zu der Frage der Automatisierung von Entscheidungen ist anderen Studien vorbehalten. Die hier vorgelegten Überlegungen sollen einen solchen Diskurs anregen.

Bisherige Studien zeigten allerdings, dass die Verantwortlichen nur selten von den Empfehlungen des Systems abweichen. Einer der Hauptgründe liegt im zeitlichen Aufwand, der für die Begründung einer abweichenden Entscheidung aufgewandt werden muss (Jedrzej, Sztandar-Sztanderska und Szymielewicz 2015; Otto 2017). Relevant ist hier zu verstehen, wann und warum Menschen von den Empfehlungen abweichen: Wenn das ADM-System nicht ihre Vorurteile teilt? Wenn das Abweichen Nutzen für den Betroffenen und keine Kosten für die Gegenseite impliziert? Wie gehen Menschen mit der Verantwortungsverschiebung um: Erlauben die jeweiligen Abhängigkeitsstrukturen vor Ort eine freie Entscheidung der Mitarbeiter, die den Prognosen eines algorithmischen Systems ggf. auch zuwiderlaufen kann, ohne dass sie persönliche Schwierigkeiten zu fürchten haben? Die reale Entscheidungsautonomie muss demnach auch in Abhängigkeit des institutionellen Rahmenbedingungen betrachtet werden: Gibt es Anreize dafür, Entscheidungen im Zweifelsfall sorgfältig zu überprüfen oder wird Selbiges eher bestraft, beispielsweise durch negative Auswirkungen auf die individuelle Leistungsbewertung?

Die Ausführung zeigt, dass die Frage der Autonomie eine hohe Relevanz hat. Die EU-Datenschutz-Grundverordnung deckt mit ihren Normen jedoch nur automatisierte Systeme, wo eine algorithmische Entscheidung eine direkte Handlung impliziert. Damit ist nur ein kleiner Teil von algorithmischen Entscheidungssystemen erfasst. Bei den anderen müssen noch Regelungen für Information, Transparenz und Widerspruchsrechte gefunden werden – sei es durch eine entsprechende Gesetzgebung für Systembetreiber oder durch die Implementierung von Verfahren, die innerhalb eines ADM-Prozesses menschliche Interventionsmöglichkeiten absichern.

3.5 Sicherheit: Wie gut ist das Entscheidungssystem gegenüber Manipulation geschützt?

Algorithmische Entscheidungssysteme werden insbesondere in der Verbindung mit selbstlernenden Algorithmen bzw. Künstlicher Intelligenz als Sicherheitsrisiko beschrieben. Ein Teil solcher Sicherheitsrisiken wurde bereits im Kontext der **adäquaten Funktionsweise** debattiert (vgl. Kapitel 2; Amodei et al. o. J.). Auch wenn Dystopien im Allgemeinen als unangebracht gelten (Ramge 2018), beschreiben Autoren wie Stephen Hawking und Elon Musk eine sich potenziell verselbstständigende, menschliche Intelligenz übersteigende Künstliche Intelligenz als elementares Risiko für die Menschheit (Future of Life Institute 2017).

Konkretere Ansatzpunkte zur **Genese von Risiken** ergeben sich laut Scherer (2016) insbesondere aus Besonderheiten der Entwicklung algorithmischer Entscheidungssysteme: Einerseits findet ein Großteil der Entwicklung von Einzelkomponenten in einem bislang unregulierten, internationalen Setting statt, das von einer unüberschaubaren Anzahl von Akteuren bestimmt werde.[8] Andererseits tritt das Potenzial algorithmischer Entscheidungssysteme erst im Zusammenwirken dynamischer Technologien zutage. Dadurch sind Risiken hinsichtlich der Funktionalität und Kontrolle gegeben. Unterschiedliche territoriale Rechtsbereiche befördern mangelnde Verantwortlichkeit. Dazu kommen potenzielle **Manipulationen**, etwa Veränderungen in der Zielorientierung, das externe Abschalten von algorithmischen Entscheidungssystemen oder die externe Veränderung der Entscheidungslogik (Amodei et al. o. J.; Future of Life Institute 2017; Russel, Dewey und Tegmark 2015).

Leider finden diese und weitere Probleme der Sicherheit von algorithmischen Entscheidungssystemen bislang wenig Eingang in politische Debatten. Wenn das der Fall ist, stehen Fragen der Standardentwicklung und Zertifizierung im Mittelpunkt. Allerdings ist häufig noch unklar, wie hoch das Sicherheitsniveau algorithmischer Entscheidungssysteme in verschiedenen Anwendungsbereichen sein soll. Auch gibt es bislang kaum Vorschläge dafür, wer Sicherheitsstandards setzen soll und wie diese überprüft werden können (Calo 2017: 14 ff.).

[8] Diese allgemeine Risikoidentifikation kann aktuell nicht illustriert werden, da es keine wissenschaftlich fundierte Übersicht über die Entwicklung einzelner Technologien durch einzelne Akteure gibt. Eine solche herzustellen, bleibt zukünftigen Untersuchungen vorbehalten.

Nimmt man die Einzelkomponenten algorithmischer Entscheidungssysteme in den Blick, lassen sich die folgenden Sicherheitsrisiken skizzieren:

- **falsche oder fehlerhafte Daten**: Dieses Risiko steigt, wenn Daten unreguliert gehandelt, kombiniert und zweckentfremdet oder zu für Betroffene unbekannten Analysezwecken eingesetzt werden (Christl 2014; 2017).
- **Entscheidungslogik**: Daten können gezielt manipuliert werden, um die Entscheidungslogik bei Formen des maschinellen Lernens zu verändern. Aktuelle Presseberichte thematisieren zusätzlich die Entwicklung von Methoden automatisierter Datenmanipulation zum Zwecke der Beeinflussung der Entscheidungslogik selbstlernender Systeme (Laskowski 2017). Ein Beispiel aus dem Medienalltag ist die Wirkung und Funktionsweise von Social Bots, die durch Verbreitung von „Fake News“ (Daten) Trends und Empfehlungssysteme von Plattformen (Entscheidungslogik) beeinflussen können.[9]
- **gesellschaftliche Einbettung**: Die Entwicklung automatisierter Entscheidungssysteme ist risikobehaftet, da sie international an unterschiedlichen Orten von unterschiedlichen Akteuren vorangetrieben wird, die verschiedenen Rechtsprechungen unterliegen (Scherer 2017): Algorithmen können beispielsweise an dem einen Ort auf Basis eines bestimmten Datensatzes trainiert, aber woanders in Software implementiert werden. Ein Beispiel dafür ist der Test einer Gesichtserkennungssoftware am Bahnhof Berlin Südkreuz. Nach Auskunft der Bundesregierung ist die durch den Systembetreiber zum Training der Algorithmen verwendete Datenbasis unbekannt. Die Informationen zum Analysesystem bleiben mit Verweis auf automatisierte Mustererkennung recht allgemein und lassen Spielraum für unterschiedliche Ziele (Deutscher Bundestag 19. Wahlperiode 2018: 7). Wie kann gewährleistet werden, dass sie im Interesse der jeweiligen Gesellschaft wirksam wird?

Welche Möglichkeiten bleiben dem Gesetzgeber darüber hinaus, Unsicherheit, Missbrauch oder Manipulation zu verhindern? Wie kann der Entstehungsprozess algorithmischer Entscheidungssysteme nachvollzogen werden? Wer ist verantwortlich für unerklärliche Fehler? Wer haftet für systemische Risiken? Und: Wer kann Fehler untersuchen und beheben? Das Thema Sicherheit wirft aktuell mehr Fragen auf, als es Antworten gibt. Hier bedarf es dringend umfassender Analysen. Zudem bestehen aktuell rechtliche Herausforderungen, Sicherheitsprobleme zu beheben, die in Kapitel 4 weiter ausgeführt werden. Dazu gehören Restriktionen adäquater Sicherheitsforschung durch IT-Sicherheitsgesetze, Urheberrechtsrecht sowie Handels- und Privatrecht.[10]

3.6 Zwischenfazit

Das Kapitel systematisiert spezielle Herausforderungen, die sich bei der Gestaltung, Anwendung und Bewertung algorithmischer Entscheidungssysteme mit Teilhaberelevanz stellen.

Im Fokus stehen zunächst die Diskriminierungsrisiken, die durch algorithmische Entscheidungen ungewollt gefestigt und befördert werden können. Die Ausführungen zeigen, dass hier nicht nur Sensibilität im Design

[9] Bislang nicht thematisiert sind Risiken, die sich aus den unterschiedlichen Analyseverfahren selbst ergeben. Unterschiedliche Suchalgorithmen haben beispielsweise Vor- und Nachteile in Abhängigkeit von der verfolgten Tiefe und Strategie, die nach Maßgabe von Zeitressourcen und Speicherkapazitäten durchgeführt werden können und sollen (Russel und Norvig 2012). Eine umfassende Risikoanalyse, die Analyseziele als auch die verschiedenen Methoden berücksichtigt und für die gesellschaftliche Debatte aufbereitet, steht aus der Perspektive der Autoren aus.

[10] Bei der Frage, wer Fehler untersuchen und beheben kann, sind die Verfügbarkeit von technischen Infrastrukturen, von Rechnerleistung sowie das Know-how von Experten von gravierender Bedeutung. Diese Aspekte werden hier nicht weiter erläutert, müssen aber beim Aufbau staatlicher Ressourcen (vgl. Kapitel 4.4) Berücksichtigung finden.

algorithmischer Systeme erforderlich ist. Vielmehr gilt es als Grundlage dafür, eine normative Debatte darüber zu führen, welches Maß an Diversität gesellschaftlich erwünscht ist.

Das gemeinwohlorientierte Design steht allerdings vor Herausforderungen, die sich aus der Entwicklung algorithmischer Entscheidungssysteme ergeben. Die Erörterung zu deren Zielsetzung und Evaluation zeigt auf, dass an verschiedenen Etappen der Systementwicklung unterschiedliche Akteure elementare Wertentscheidungen treffen müssen. Dies erschwert die Durchsetzung normativer Leitlinien. Die Realität ist durch vielfältige Verantwortlichkeiten charakterisiert, denen Rechnung zu tragen ist.

Probleme der Überprüfbarkeit ergeben sich insbesondere aus der technologischen Entwicklung, die immer komplexere und dynamischere Entscheidungssysteme möglich macht. Diese finden sich zwar aktuell vornehmlich im privaten Sektor, der über entsprechende Datenakkumulations- und Analysekapazitäten verfügt. Am Beispiel der vorhersagebasierten Polizeiarbeit (Predictive Policing) konnte jedoch gezeigt werden, dass diese komplexen, dynamischen Entscheidungssysteme auch Eingang in den öffentlichen Sektor halten. Damit müssen Lösungen für komplexe Systeme in kooperativen Regulierungsarrangements gefunden werden.

Darüber hinaus stellt sich die Frage der Automatisierung algorithmischer Entscheidungssysteme auch bei der Anwendung algorithmischer Entscheidungen. Die prototypische Unterscheidung in Systeme, die der Entscheidungsunterstützung dienen, und solchen, die unmittelbare Wirkung auf Individuen entfalten, dürfte in der Praxis zunehmend verwischen. Denn auch Empfehlungssysteme sind auf interventionsfreie Abläufe optimiert, welche die Prämisse menschlicher Interventionsoptionen unter Umständen unterminieren. Daher müssen Regelungen für Information, Transparenz und Widerspruchsrecht auch für teilautomatisierte Assistenzsysteme gefunden werden.

Die Entwicklung und Komplexität von algorithmischen Entscheidungssystemen, die Technologien Künstlicher Intelligenz beinhalten, birgt besondere Sicherheitsrisiken. Diese ergeben sich zum Teil aus der über verschiedene Bereiche und Akteure verteilten Entwicklung, zum Teil aus den Manipulationsmöglichkeiten einzelner Komponenten. Neben Design- und Anwendungsfehlern erhöhen diese systemischen Risiken die Relevanz einer adäquaten Überprüfbarkeit algorithmischer Entscheidungssysteme.

Das Kapitel vertieft damit das Verständnis für die Besonderheiten und Herausforderungen bei der gemeinwohlorientierten Entwicklung und Gestaltung algorithmischer Entscheidungssysteme. Darauf aufbauend werden im Folgenden mögliche Lösungsansätze skizziert und systematisiert.

4 Was man tun kann: Panorama der Lösungsvorschläge

Im Folgenden werden Lösungsansätze skizziert und systematisiert, die in die internationale Debatte zur Gewährleistung einer gemeinwohlorientierten Entwicklung algorithmischer Entscheidungssysteme Eingang gefunden haben. Im Fokus stehen vier Handlungsfelder, die das Kapitel zu gesellschaftlichen Anforderungen (vgl. Kapitel 2.2.) herausgearbeitet hat. Es sind:

- Zielsetzung algorithmischer Systeme
- Umsetzung der Ziele im Einsatz
- Vielfalt der Systeme, Ziele und Betreiber
- übergreifende Rahmenbedingungen für den Einsatz

Die vier Handlungsfelder liegen auf unterschiedlichen Analyseebenen:

- Zielsetzung und Umsetzung müssen für jedes einzelne algorithmische System diskutiert und überprüft werden.
- Die Vielfalt der Systeme, Ziele und Betreiber kann nur auf Ebene der Gesamtheit aller Systeme überprüft und gesichert werden.
- Übergreifende Rahmenbedingungen für den teilhabeförderlichen Einsatz algorithmischer Systeme, wie zum Beispiel Kompetenz bei Betroffenen, Anwendern sowie staatliche Regulierungskompetenz, liegen quer zu den vier oben skizierten Handlungsfeldern und wirken auch auf alle vier.

Die vier Handlungsfelder sind prototypisch zugeschnitten. Einzelne Lösungsansätze können durchaus mehrere Felder betreffen. So haben zum Beispiel zivilgesellschaftliche Watchdog-Organisationen einerseits die Möglichkeit, die Umsetzung von Zielen in bestimmten Systeme prüfen. Andererseits nehmen sie eine starke Rolle ein in der Debatte um die Angemessenheit der Optimierungsziele und der Beförderung des gesellschaftlichen Diskurses dazu. Bei solchen möglichen Überlappungen haben wir diese im jeweiligen Steckbrief vermerkt und uns in der folgenden Struktur bei der Zuordnung für das Handlungsfeld entschieden, in dem wir das größere Wirkungspotenzial sehen. Die einzelnen Ideen in den Handlungsfeldern sind in aufsteigender Reihenfolge nach dem bisherigen Konkretisierungsgrad geordnet.

Das folgende Lösungspanorama zeigt ganz deutlich: Es gibt nicht die eine Lösung für alle Herausforderungen algorithmischer Entscheidungsfindung, sondern ein Spektrum an Ansätzen, welche dazu beitragen können, algorithmische Systeme in den Dienst von Mensch und Gesellschaft zu stellen. Viele Ideen sind noch zu konkretisieren, damit überhaupt Konzepte zur Debatte stehen. Das Arbeitspapier soll Anfang und Anstoß für diese Aufgaben sein.

4.1 Zielsetzung algorithmischer Systeme auf gesellschaftliche Angemessenheit prüfen

Welche Optimierungsziele gesellschaftlich sinnvoll sind, lässt sich nicht allgemein für alle algorithmischen Systeme festlegen. Definition und Priorisierung gesellschaftlicher Ziele sind ein dynamischer Prozess. Jedes neue algorithmische System ist Anlass und Anstoß, diesen Prozess weiterzutreiben.

> *„From our perspective, addressing the ethical implications of AI poses a dilemma because questions of ethics are about processing and evaluating risks and benefits or acceptable trade-offs in specific circumstances. The area of ethics should not be thought of as prescriptive, but rather as requiring processes for assessing multiple perspectives and outcomes"* (Data & Society 2017: 1).

Wie kann die Entwicklung von algorithmischen Systemen gestaltet sein, um in teilhaberelevanten Anwendungsfällen eine sinnvolle Zielsetzung zu ermöglichen? Das ist die Kernfrage, welche die im Folgenden skizzierten Lösungsansätze beantworten. Wir fokussieren hier die Gestaltung des Zielsetzungsprozesses, nicht die Festlegung einer klaren Hierarchie von in jedem Anwendungsfall zu verfolgenden Zielen.

4.1.1 Interessen, Stakeholder und Optimierungsziele dokumentieren

Steckbrief: Interessenmatrix

Kerngedanke: Unterschiedliche Optimierungsziele sowie damit verbundene Interessen und Stakeholdergruppen erkennen, in Beziehung setzen, dokumentieren

Handlungsfeld: Optimierungsziele auf Angemessenheit prüfen

Stakeholder: Entwickler, Systembetreiber, alle potenziell vom ADM-System betroffenen Stakeholder

Durchsetzende Akteure: Entwickler, Systembetreiber, Nichtregierungsorganisationen, Fachverbände, Staat, Standardisierungsinstitutionen

Instrumente: Prozessstandard, Dokumentationsstandard

Status: Idee

Wer ein algorithmisches System entwickelt, muss Ziele priorisieren. Nehmen wir eine einfach erscheinende Aufgabe: Eine Software soll Patienten auf die Betten in verschiedenen Stationen eines Krankenhauses verteilen. Ohne tiefergehende Recherche ist es naheliegend, dass diese Software auf unterschiedliche Ziele und damit verbundene Interessen hin optimiert werden kann. Wie zum Beispiel:

- die bestmögliche Auslastung des Krankhauses
- die größtmögliche Abrechenbarkeit der Leistungen bei Versicherungen
- die höchstmögliche Versorgungsqualität für die Patienten
- das Renommee einer Institution (vgl. Cohen et al. 2014)

Solche Interessen sind bei der Entwicklung eines algorithmischen Systems darzustellen und zu dokumentieren. Wie Optimierungsziele, Interessen und unterschiedliche Stakeholdergruppen zusammenhängen, kann in einer Art **Interessenmatrix** dargestellt werden, die eine Grundlage und Teil einer Folgenabschätzung oder Verträglichkeitsprüfung sein kann (vgl. Kapitel 4.1.3).[11] Indem die beteiligten Akteure alle potenziell bei Entwicklung und Einsatz eines ADM-Systems wirksamen und betroffenen Interessen dokumentieren, schaffen sie die Grundlage dafür, Stakeholder zu identifizieren und einzubinden.

Eine Konkretisierung der Idee einer Verträglichkeitsprüfung sollte auch folgende Fragen adressieren:

- Wie lässt sich eine bewusste oder aus Unkenntnis einseitige Darstellung der Interessen verhindern?
- Welche Akteure kommen für die Standardisierung und Qualitätssicherung beim der Erstellung der Interessenmatrix infrage?
- Können rechtliche Instrumente solche Werkzeuge zur Prüfung der Verträglichkeit fördern (z. B. Belegpflicht, dass die Interessenmatrix nach einem anerkannten Verfahren erstellt und veröffentlicht wurde)?

[11] Die Idee der Dokumentation in einer Interessenmatrix ist im Rahmen eines Workshops entstanden; die Autoren danken Udo Seelmeyer und Andreas Dewes für die anregende Diskussion.

4.1.2 Betroffene über den ADM-Einsatz informieren

Steckbrief: Informations- und Transparenzpflichten

Kerngedanke: Betroffene über den Einsatz und die Zielstellung von ADM informieren

Handlungsfelder: Optimierungsziele auf Angemessenheit prüfen, Umsetzung prüfen

Stakeholder: Systembetreiber

Durchsetzende Akteure: Systembetreiber, Staat

Instrumente: Gesetz, Selbstverpflichtung

Status: eine erste Implementierung in der DSGVO-EU, erste ausgearbeitete Idee (Anwendungserläuterung/„counterfactual explanations“)

Grundvoraussetzung für die Gewährleistung gesellschaftlicher Angemessenheit bei der algorithmischen Entscheidungsfindung stellt die Kenntnis über den Einsatz derartiger Verfahren dar. Nur wo bekannt ist, dass algorithmische Systeme im Einsatz sind, haben Betroffene und ihre Interessenvertreter die Gelegenheit, die Wirkung, die Ziele und die Umsetzung der eingesetzten Software zu untersuchen.

Die Europäische Datenschutz-Grundverordnung (DSGVO-EU) sieht im Kapitel 3, Art. 13 bis 15 in Verbindung mit Art. 22 **Informationspflichten seitens der Systembetreiber** für automatisierte Entscheidungen vor (zu teilautomatisierten Entscheidungs- bzw. Assistenzsystemen siehe unten). Das betrifft einerseits erhobene und (vgl. Cohen et al. 2014) verarbeitete Daten, insofern diese Eingang in algorithmische Prozesse finden. Andererseits umfasst die Verordnung aussagekräftige Informationen über die a) involvierte Logik, b) die Tragweite sowie c) die angestrebten Auswirkungen automatisierter Entscheidungen – zumindest in den Fällen, wo die Entscheidung gegenüber den Betroffenen rechtliche Wirkung entfaltet oder sie in ähnlicher Weise erheblich beeinträchtigt (zu den Ausnahmen siehe a. a. O.).

Ähnlich fordert der Europarat in einer Studie zu den Menschenrechtsdimensionen automatisierter Entscheidungsverfahren eine **effektive Transparenz** über:

- Ziele der algorithmischen Entscheidungsfindung
- Variablen im Sinne der Modellierung
- Informationen über Methoden (Trainingsdaten, statistische Kennzahlen sowie die Menge und Art der automatisierten Entscheidungen zugrunde liegenden Daten, vgl. Council of Europe – Committee of experts on internet intermediaries 2017)

Citron (2008: 1281 ff.) hebt hervor, dass solche Auskunftspflichten grundlegend dafür sind, dass rechtsstaatliche Verfahren („due process“) in prozeduraler und inhaltlicher Hinsicht gewahrt werden können.

Die Forderung nach **Transparenz- und Auskunftspflichten** im Kontext von Zielen und Entscheidungslogik sowie Methodik ist den Ansätzen zur Überprüfung von Algorithmen inhärent. Zentral ist zudem die Kenntnis der Datengrundlage (Diakopoulos 2016; Zweig 2016).

Solche Forderungen bleiben allerdings oft recht allgemein: Die Normen der EU-Datenschutz-Grundverordnung beziehen sich ausschließlich auf automatisierte Entscheidungen, welche den individuellen Handlungsspielraum beeinflussen. Daneben bestehen aber auch automatisierte Entscheidungssysteme, bei denen von einer Beeinflussung unserer Wahrnehmung ausgegangen werden muss, wie zum Beispiel soziale Netzwerke (Google, Facebook, Twitter) oder Plattformen, die dem Onlineversandhandel (Amazon) oder der Verteilung anderer Güter oder Dienstleistungen dienen (Otto 2017: 18 ff.). Wenngleich die potenzielle Beeinflussung der Wahrnehmung zunächst keine rechtliche Wirkung entfaltet, stellt sie nach Lenk (2016) durch die Beeinflussung der Informationsräume von Individuen und Organisationen doch eine indirekte Verhaltenssteuerung dar und kann gesellschaftlich bedenklich sein. Ein Beispiel dafür stellen die Kaufempfehlungen Amazons für Sprengstoffzutaten dar, die laut Medienberichten die Vorbereitungen von Terrorverdächtigen substanziell unterstützt haben sollen (Beuth 2017).

Folgerichtig fordern Experten auch eine gesetzliche Verpflichtung zur **Transparenz von Algorithmen in sozialen Netzwerken** (Mittelstadt 2016; Pasquale 2010).[12]

Darüber hinaus müssen Lösungen im Bereich Transparenz- und Auskunftspflichten für algorithmische Empfehlungssysteme gefunden werden (vgl. Kapitel 3.4). Denn wenngleich Empfehlungssysteme in Einsatzbereichen wie dem Gesundheits- oder Sicherheitsbereich menschliche Interventionsmöglichkeiten implizieren, zielen sie auf einen interventionsfreien Ablauf und haben damit im Regelfall eine ähnliche Wirkung wie automatisierte Entscheidungen. Folgerichtig macht etwa die Stadt New York bei der Regulierung algorithmischer Entscheidungssysteme in öffentlicher Hand keinen Unterschied zwischen automatisierter Entscheidung und Empfehlung. In beiden Fällen sollen Entscheidungen überprüfbar und nachvollziehbar sein (The New York City Council 2018).

Eine Herausforderung bei der Umsetzung von Transparenz- und Auskunftsrechten bleibt allerdings die Dynamik und Komplexität neuerer Systeme: Wie können Aussagen zu involvierter Logik, zur Tragweite und zu angestrebten Auswirkungen getroffen werden, wenn diese mitunter regelmäßig variieren? Der Suchalgorithmus von Google stellt ein Beispiel dafür dar, dass Transparenz voraussetzungsvoll ist: Ein komplexes System aus automatisierter Analyse von Webseiten, Indexierung und Ranking mit über 200 Variablen wird hunderte Male pro Jahr verändert (Pasquale 2016).

Im Allgemeinen ist anzunehmen: Das Ineinandergreifen von einzelnen Softwarekomponenten, die von unterschiedlichen Programmierern mit begrenztem Überblick erstellt wurden, bietet seit Jahrzehnten die Grundlage für die Unwägbarkeit algorithmischer Prozesse (Passig 2017). Nun finden selbstlernende Algorithmen Eingang in algorithmische Entscheidungen (vgl. Kapitel 2.1.2), deren Ergebnisse unabhängig von der Zielsetzung schwer zu determinieren und zu erklären sind. Sie finden Anwendung in immer mehr Lebensbereichen. Welche Möglichkeiten bieten sich der Gesellschaft, die Entwicklung und Resultate algorithmischer Entscheidungen zu verstehen und mitzugestalten?

Wissenschaftler des Oxford Internet Institute haben hierzu einen instruktiven Vorschlag gemacht, der als **Counterfactual Explanations** – eine Art **Anwendungserläuterung** für algorithmische Entscheidungssysteme – debattiert wird. Die Idee dahinter ist folgende:

Es gibt Gründe, die es schwierig bis unmöglich machen, algorithmische Entscheidungen zu erklären. Dazu zählen technische Probleme im Kontext komplexer, selbstlernender Systeme, die im Folgenden als **Black Box** bezeichnet werden (vgl. Kapitel 3.3). Aber auch rechtliche Aspekte können dazu führen, dass eine generell zu unterstützende Transparenz im Einzelfall nicht akzeptabel ist. Neben Geschäftsgeheimnissen bieten der Datenschutz von Dritten sowie die Gefahr der Manipulation des Entscheidungssystems Anlass, die von der Öffentlichkeit gestellten Forderungen nach Transparenz abzuwägen. Doch auch in Fällen, in denen Transparenz schwierig ist, sollten Betroffene einerseits die Möglichkeit haben, algorithmische Entscheidungen zu verstehen und anzufechten. Andererseits sollten ihnen Orientierungshilfen vorgehalten werden, wie sie in Zukunft auf algorithmische Entscheidungen, beispielsweise durch Änderung ihres Verhaltens, Einfluss nehmen können.

> *„These counterfactual explanations describe the smallest change to the world that would obtain a desirable outcome, or to arrive at a ‚close possible world'"* (Wachter, Mittelstadt und Russell 2017: 1).

[12] Auch wenn die Regulierung kommerzieller Plattformen nur bedingt (zum Zwecke der Illustrierung und Abgrenzung) Gegenstand der Lösungsvorschläge dieses Arbeitspapiers ist, sei angemerkt, dass eine Transparenz die Voraussetzung darstellen würde für weitergehende Forderungen nach Kontrolle oder Gestaltung der Empfehlungssysteme. Eine Sonderform der Gestaltung würde die Einführung von Auswahloptionen für Kunden darstellen, etwa die Möglichkeit, nach Relevanz sortierte Nachrichteninhalte zugunsten chronologischer Anzeige auszuschalten.

Im einfachsten Fall, etwa bei der Bewilligung von Krediten, könnte eine Anwendungserläuterung eine Auskunft dazu geben, wie hoch das Jahreseinkommen eines Antragstellers sein müsste, damit sein abgelehnter Kreditantrag bewilligt wird. Interessant wird diese Methode, wenn viele Variablen oder komplette Variablensets vorliegen. In diesem Fall würde die Analyse der einer algorithmischen Entscheidungsfindung zugrunde liegenden Variablen verschiedene Erklärungen für das Resultat bieten können. Für das von einer Entscheidung betroffene Individuum wäre voraussichtlich die Erklärung am hilfreichsten, welche die individuell nächstmögliche Veränderung fokussiert, so die Autoren (a. a. O.). Schlussfolgernd ist das Ergebnis dieser Methode eine algorithmisch generierte Erklärung algorithmischer Entscheidungsverfahren.

Neu an diesem Ansatz ist die innovative Transformation von Recht in Code zum Zwecke der Verbraucheraufklärung. Die mathematische Methode hat aber auch Tücken: Ein algorithmisches Entscheidungssystem soll durch ein algorithmisches Erklärungssystem ergänzt werden. Es generiert damit prospektiv die gleichen Fragen wie das Ursprungsproblem, etwa:

- **Nachvollziehbarkeit**: Können Anwendungserläuterungen nachvollzogen werden? Liegt ihr Quellcode vor, der eine Prüfung ermöglicht, oder unterliegt er der Geheimhaltung?
- **Kontrolle**: Funktionieren Anwendungserläuterungen fehlerfrei und angemessen? Inwiefern sind sie abhängig von der Verfügbarkeit einer adäquaten Datenbasis?
- **Sicherheit**: Kann eine Zuverlässigkeit algorithmischer Erklärungen gewährleistet werden, wenn selbstlernende Algorithmen zum Einsatz kommen, die auf Basis von Wahrscheinlichkeiten arbeiten anstelle von Exaktheit?

Die Idee von Anwendungserläuterungen stellt einen konstruktiven Ausgangspunkt für verbraucherfreundliche Erläuterungen algorithmischer Prozesse in bestimmten Anwendungsbereichen dar, der ausbaufähig zu sein scheint und fortentwickelt werden muss, um Vertrauen zu schaffen. Zudem scheint es angemessen, potenzielle Anwendungsfelder näher zu bestimmen und die jeweilige Teilhabe- und Risikorelevanz zu berücksichtigen. Eine Gefahr dürfte darin bestehen, dass viele potenzielle Fehlerquellen nicht identifiziert werden können. Dazu zählen insbesondere:

- die Qualität der Daten in einer bestimmten Datenbank,
- Verzerrungen der Datenstruktur,
- Über- oder Unterrepräsentation bestimmter Profile und
- die Entscheidungsarchitektur des ADM-Systems (und viele Fehler liegen eben darin).

Diese Problembereiche liegen im Design- und Anwendungsbereich, der ggf. neue Kontrollsysteme und -verfahren erfordert.

Wie Informations- und Transparenzpflichten sinnvoll gegenüber den Betroffenen umgesetzt werden können, wird in der Debatte über algorithmische Entscheidungsverfahren bislang vernachlässigt. Ein Ansatzunkt ist die Idee des Datenbriefs vom Chaos Computer Club: Sie sieht vor, dass Betroffene von Firmen, Behörden und Institutionen regelmäßig über ihre gespeicherten personenbezogenen Daten informiert werden sollen. Der Vorschlag erstreckt sich auch auf Daten, die durch Verarbeitung und Kombination mit anderen Daten erstellt wurden, wie z. B. Profile, Annahmen über Präferenzen oder interne Kundenklassenzuordnungen (frankro 2010). Die Möglichkeit des Verbandsklagerechts wird in Kapitel 4.2.4 erläutert.

4.1.3 Die erwarteten Folgen und Auswirkungen reflektieren und dokumentieren

Steckbrief: Algorithmenfolgeabschätzungen und Verträglichkeitsprüfungen

Kerngedanke: Der Öffentlichkeit einen Überblick über den Einsatz und die Zielstellung von ADM geben, als Grundlage für weitere Maßnahmen (Algorithmen-TÜV, Stakeholderpartizipation, Prüfung)

Handlungsfelder: Optimierungsziele auf Angemessenheit prüfen, Umsetzung prüfen

Stakeholder: Entscheider, Systembetreiber

Durchsetzende Akteure: Systembetreiber, Fachverbände, Staat

Instrumente: Verpflichtung, Selbstverpflichtung, Prozessstandard, Dokumentationsstandard

Status: Idee, frühe Prototyen (FAT/ML), erprobte Vorbilder aus anderen Bereichen (z. B. Bauwirtschaft)

Transparenz- und Auskunftspflichten sollen Betroffenen einen Überblick über die Grundlagen einer sie betreffenden Entscheidung geben und darauf aufbauend die Möglichkeit ihrer Anfechtung und Veränderung. Wie aber ist es um diesen Überblick für die Gesellschaft bestellt?

Folgeabschätzungen bzw. **Verträglichkeitsprüfungen** für algorithmische Entscheidungssysteme können eine Grundlage sein, der Öffentlichkeit einen Überblick über eingesetzte Systeme zu bieten. Denn nur mit Überblick könne dem Einsatz fehlerhafter, diskriminierender oder schädlicher Systeme Einhalt geboten werden kann, so der Informatiker Ben Shneiderman (2016: 13539). Er schlägt vor, Systembetreiber zum Bereitstellen einer Folgenabschätzung zu verpflichten. Diese könnte dann Grundlage für ein unabhängiges Zulassungsverfahren („independent oversight review“) sein.

Diese Vorschläge orientieren sich an den in der Europäischen Union und den USA bekannten Umweltverträglichkeitsprüfungen bzw. -erklärungen, die zum Beispiel bei Bauvorhaben Informationen über betroffene Stakeholdergruppen und potenzielle Folgen geben. Eine **Algorithmenverträglichkeitsprüfung** bzw. **-erklärung** („algorithm impact statement“) könnte der Öffentlichkeit einen Überblick geben über die

- Ziele eines algorithmischen Entscheidungssystems,
- Qualität des Dateninput und
- erwartete Ergebnisse (a. a. O.).

So wären beispielsweise Abweichungen von einer intendierten Funktionsweise identifizierbar. In der Diskussion in Deutschland wird eine ähnliche Idee unter dem Schlagwort „Beipackzettel“ diskutiert – ein Dokument, das „Einsatzgebiet, Modellannahmen und gesellschaftliche Nebenwirkungen“ eines algorithmischen Systems benennt (Zweig 2016). Eine Konkretisierung der Idee Verträglichkeitsprüfung sollte auch diese Themen adressieren:

- erwartete individuelle Nebenwirkungen
- erwartete Qualität (z. B. Fehlerquoten)
- sinnvolle Verfahren zur Qualitätssicherung im Einsatz (z. B. Vergleich Trainings- und Realdaten bei automatisch lernenden Systemen)

In eine ähnliche Richtung geht die Idee einer sogenannten **Diskriminierungsprüfung** bzw**. -erklärung** („discrimination impact assessment“), wie sie der Rechtswissenschaftler Andrew Selbst (2016) für die vorhersagebasierte Polizeiarbeit („predictive policing“) vorschlägt: Hier stehen Effektivität und potenzielle Diskriminierungseffekte algorithmischer Entscheidungssysteme im Mittelpunkt der Betrachtung. Der Vorschlag beinhaltet den Vergleich unterschiedlicher Algorithmen und Modelle im gleichen Anwendungsbereich. Er würde der Öffentlichkeit eine Möglichkeit für die Auswahl, Mitgestaltung oder Entwicklung algorithmischer Entscheidungssysteme eröffnen. Chancen und Risiken der vorhersagebasierten Technologien könnten langfristig debattiert werden. Zudem würde das Vertrauen in die Arbeit von Polizei- und Sicherheitsdiensten gestärkt.

Die Forderung nach einer Art **sozialer Verträglichkeitsprüfung algorithmischer Entscheidungssysteme** wird auch von dem Expertennetzwerk **Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)** geäußert. Die Autoren nehmen das Entscheidungssystem und seinen Entwicklungs- und Anwendungskontext in den Blick. Sie fordern im Rahmen einer solchen gesellschaftlichen Folgenabschätzung Informationen seitens der Betreiber auch zu Haftungsfragen, Informations-, Transparenz- und Widerspruchspflichten sowie Algorithmenauditing (FAT/ML 2016).

Zum Erstellen einer solchen Abschätzung in Form einer **Risikoprognose** können die Gestalter eines algorithmischen Systems in bestimmten Einsatzgebieten auch gesetzlich verpflichtet werden. Diesen Vorschlag führt Martini aus:

> *„Wer in seine Softwareanwendung Algorithmen implementiert, die das Potenzial erheblicher Persönlichkeits-, insbesondere Diskriminierungsrisiken bergen, sollte grundsätzlich eine Risikoprognose erstellen müssen: Er muss dann analysieren und offenlegen, inwieweit das digitale System grundrechtlich geschützte Güter gefährdet und welche technischen, organisatorischen und rechtlichen Schutzmechanismen er vorsieht, um Rechtsverletzungen zu vermeiden“* (Martini 2017: 1022).

Systembetreiber sind hierbei angesprochen, entsprechende Informationen bereitzustellen. Verschiedene Modelle zwischen Selbst- und Ko-Regulierung sind denkbar, etwa die gesetzliche Verpflichtung von Systembetreibern wie auch deren Selbstverpflichtung auf bestimmte Standards.

4.1.4 Partizipation aller Stakeholder bei Entwicklung und Anwendung sichern

Steckbrief: Gutachtergremien für Stakeholderpartizipation

Kerngedanke: Verfahren der Stakeholderpartizipation in der Entwicklung, Implementierung und Evaluation von ADM institutionalisieren

Handlungsfeld: Optimierungsziele auf Angemessenheit prüfen

Stakeholder: Entwickler, Systembetreiber, Anwender, alle potenziell vom ADM-System Betroffenen

Durchsetzende Akteure: Entwickler, Systembetreiber, Fachverbände, Standardisierungsinstitutionen

Instrumente: Prozessstandard, Beteiligungsformat

Status: Idee, mögliche Vorbilder in anderen Bereichen (Gutachtergremien/„IT review boards“)

Die Entwicklung und der Einsatz algorithmischer Entscheidungssysteme beinhalten von Beginn an Wertentscheidungen, die in das Systemdesign einfließen.

Wie das bereits erörterte Beispiel zur personalisierten Diagnose und Behandlung (vgl. Kapitel 3.2) zeigte, können hier **widersprüchliche Interessen** aufeinandertreffen. Daher fordert die Rechtswissenschaftlerin Danielle Keats Citron (2008: 1288 ff.), betroffene Stakeholder in den Entwicklungsprozess teilhaberelevanter algorithmischer Entscheidungssysteme einzubeziehen. Denn die Entwicklung algorithmischer Entscheidungssysteme beinhalte ein **Übersetzen von politischen und gesellschaftlichen Zielen und Programmen in Code.** Damit komme sie einer Art Gesetzgebung gleich. Diese müsse einerseits eine Veröffentlichung der gesetzten Regeln, andererseits Möglichkeiten ihrer öffentlichen Debatte absichern. Anderenfalls werde der demokratische Prozess unterwandert.

Diese Forderung scheint elementar in **Anwendungsbereichen, die teilhaberelevant sind**. Positive Erfahrungen ergaben sich in der Vergangenheit im US-amerikanischen Raum in Bereichen wie der sozialen Sicherung. Eine konkrete Möglichkeit, Partizipation zu institutionalisieren, besteht in der **Bildung von Gutachterkreisen** („information technology review boards“), die Experten und der Öffentlichkeit Partizipation in der Systementwicklung, Implementierung, Anwendung und Evaluation bieten (a. a. O.: 1312).

Vorbilder für adäquate **Verfahren der Stakeholderpartizipation** finden sich auch im Bereich der Bio- und Medizinethik. Im Kontext von Genforschung wurden alternative Lösungen für ethische Probleme auf verschiedenen

gesellschaftlichen Ebenen gefunden. Die **Einbindung von Patientenvertretungen** in Organisationen, die mit der Entwicklung und Anwendung algorithmischer Entscheidungsfindung in der Medizin betraut sind, stellt eine Lösungsmöglichkeit dar, so Cohen et al. (2014). Positive Erfahrungen liegen für Nutzung und Verwertung von Gendatenbanken vor, die ein **Treuhänder** (Individuum oder eine Gruppe von Betroffenen) überschaut. Zu klären ist, welche Rechte ein Treuhänder im Falle von Missbrauch oder Problemen bei algorithmischen Entscheidungssystemen hätte. Alternativen zu einem solchen Modell bestünden in der Einrichtung von **kommunalen Gutachtergremien**, wo Stakeholder wie Patienten, Ärzte, Krankenhäuser und Data Scientists gemeinsame Empfehlungen entwickeln (a. a. O.).

Die Übersicht zeigt, dass eine **Vielzahl an Institutionalisierungsmöglichkeiten** für die Partizipation von Stakeholdern möglich ist. Wie bereits im Kontext der Folgeabschätzungen (vgl. Kapitel 4.1.3) erörtert, ist die Entwicklung von algorithmischen Entscheidungssystemen, ihre konkrete Implementierung und Anwendung sowie ihre Evaluation zu berücksichtigen.

Eine Konkretisierung der Idee muss auch die Fragen adressieren, wie die potenziell betroffene breite Öffentlichkeit in diese Prozesse effektiv eingebunden werden kann. Welche Akteure können wirksam als Übersetzer der fachlichen Fragen für ein breites Publikum fungieren? Wie können gestalterische Details, deren Diskussion Fachwissen erfordert, in hier relevante gesellschaftliche Fragen übersetz werden?

Insgesamt fallen solche Formen der Regulierungen in den Bereich von **Selbst- und Ko-Regulierung**.

4.1.5 Professionsethik etablieren

Steckbrief: Entwicklung professionsethischer Kodizes und Institutionen

Kerngedanke: Prozessbezogene Qualitätsstandards für die Gestaltung algorithmischer Systeme kodifizieren, um Mindestmaße zum Beispiel an Sorgfalt, Erklärbarkeit und Folgenabschätzung zu sichern

Handlungsfelder: Zielsetzung, Umsetzung

Stakeholder: Entwickler, Unternehmen, Wissenschaft, Hochschulen, Berufsverbände, Nichtregierungsorganisationen

Durchsetzende Akteure: berufsständische Organisationen

Instrument: Selbstverpflichtung

Status: Ideen, erste Konkretisierungen, mögliche Vorbilder in anderen Bereichen (Medizin, medizinzische Forschung, Journalismus, soziale Arbeit, Psychologie)

Eine Reihe von Professionen haben Prinzipien für Urteile über das Wohl von Menschen erarbeitet und Institutionen etabliert, die konkrete Fälle an diesen Prinzipien messen und beurteilen. Die vielleicht bekannteste Grundlage einer solchen Professionsethik ist der Hippokratische Eid. Dieser ist heute aktualisiert als Genfer Deklaration des Weltärztebundes noch immer beispielsweise der Musterberufsordnung für die deutschen Ärzte vorangestellt (Bundesärztekammer 2015: 2). In der Diskussion über die Bewertung gesellschaftlicher Angemessenheit algorithmischer Systeme empfehlen viele Experten Äquivalente zu professionsethischen Standards und Institutionen, wie sie etwa in medizinischer Forschung existieren. Beispielhaft für solche Empfehlungen aus dem zivilgesellschaftlichen Bereich ist die des AI Now Institute:

> *„Ethical codes meant to steer the AI field should be accompanied by strong oversight and accountability mechanisms. More work is needed on how to substantively connect high level ethical principles and guidelines for best practices to everyday development processes, promotion and product release cycles“* (Campolo et al. 2017: 2).

Für die Konkretisierung und Umsetzung sind konzeptionelle Fragen auf drei Themenfeldern zu beantworten:

1. Wer ist Adressat der Professionsethik?
2. Was steht drin?

3. Wie erlangen die Prinzipien Anerkennung und wie werden sie konkretisiert und aktualisiert?

1. Wer ist Adressat der Professionsethik? An der Entwicklung von Algorithmen, dem Erheben von Daten, dem Entwickeln von Modellen, der Implementierung sowie dem Einsatz und der Evaluation von einzelnen Systemen sind unterschiedliche Berufsgruppen beteiligt. Data Scientists sind zwar an vielen Prozessschritten beteiligt (Zweig 2018), aber nicht unbedingt an allen. Was ist etwa mit Produktmanagern oder Statistikern? Adressiert eine Professionsethik alle am Entwicklungs- und Einbettungsprozess Beteiligten, steht man vor der Herausforderung, die Aktualisierung, Verbindlichkeit und Anerkennung der Prinzipien ohne die gemeinsame Basis etwa eines Berufsverbandes zu sichern. Hinzu kommt: Abstrakte Entwürfe, die sich an alle an der Entwicklung und Implementierung algorithmischer Systeme Beteiligten richten, dürften schwieriger zu etablieren sein als Prinzipien in einer Profession mit bestehenden Berufsverbänden und Ausbildungswegen. Je schwächer die Selbstwahrnehmung als Profession in einem Bereich ist, desto schwieriger wird es, eine Bereichsethik als Professionsethik an professionelles Selbstverständnis zu knüpfen. Wie lässt sich dennoch Verbindlichkeit erreichen? Das sind Umsetzungsfragen, die bei einer Konkretisierung adressiert werden müssen. Die bislang vorliegenden Vorschläge enthalten dazu keine Anregungen.

2. Was steht drin? Algorithmische Systeme können in vielen unterschiedlichen Sektoren zum Einsatz kommen. Zum Beispiel in der medizinischen Diagnostik, in der Justiz, in der Personalauswahl. Diese Einsatzbreite ist eine besondere Herausforderung für eine Professionsethik, weil zum Beispiel Optimierungsziele und Umsetzungsfolgen des Systemoutputs bei algorithmischen Prozessen gar nicht von Algorithmikern allein bestimmt oder gar überschaut werden. Das ist in anderen Professionen wie Medizin, Journalismus oder sozialer Arbeit anders gelagert.

Ob die Optimierungsziele gesellschaftlich angemessen sind, lässt sich nicht allein auf Grundlage der Gestaltung des Systems und der Datenauswahl diskutieren. Welche Folgen hat der Einsatz für den einzelnen Bewerteten? Welche Folgen hat der Einsatz absehbar auf kollektive Güter? Welche Alternativen bestehen? Diese Fragen können bei gesellschaftlich relevanten Systemen nicht allein die Data Scientists, Produktmanager, Implementierer und andere an der Entwicklung Beteiligte beantworten. Hier braucht es viele Anknüpfungspunkte und Instrumente, um einen Diskurs in Gang zu bringen und zu ermöglichen. Dieses zu ermöglichen, ist ein Ansatz für eine Professionsethik: Wo algorithmische Systeme gesellschaftliche Teilhabe beeinflussen, müssen ihre Ziele, ihr Design und ihre Wirkung der gesellschaftlichen Kontrolle und Willensbildung unterliegen. Diese Rückkoppelung kann auch ein Treiber für die Entwicklung von Ethikkodizes sein:

> *„Es war letztlich der zivilgesellschaftliche Diskurs, welcher die bioethische Suche nach Normen in der Medizin und ihrer Begründung in medizinische Lehre, Praxis, Forschung und Institutionen einziehen ließ, zum Beispiel in der Gestalt klinischer Ethikkomitees“* (Frick SJ 2018: 103).

Diese Willensbildung zu ermöglichen, ist die ethische Verantwortung der an der Entwicklung der Systeme Beteiligten. Wir nennen diesen Ansatz im Folgenden **prozessbezogene Professionsethik**.

In diese Richtung gehen Vorschläge wie die „Principles for Accountable Algorithms“ der Initiative Fairness, Accountability, and Transparency in Machine Learning (FAT/ML). Die Autoren formulieren dieses Ziel einer prozessbezogenen Professionsethik:

> *„(…) to help developers and product managers design and implement algorithmic systems in publicly accountable ways. Accountability in this context includes an obligation to report, explain, or justify algorithmic decision-making as well as mitigate any negative social impacts or potential harms”* (FAT/ML 2016)

Die Autoren postulieren fünf Prinzipien für die Gestaltung algorithmischer Systeme unter dieser Maxime (im Folgenden aus dem englischen Original paraphrasiert):

- **Verantwortlichkeit und Prozeduralisierung**: Instanzen für Beschwerden und Berufung schaffen und öffentlich sichtbar machen.
- **Erklärbarkeit**: Sicherstellen, dass Entscheidungen Endanwendern und anderen Stakeholdern in nicht technischer Terminologie erklärt werden können.
- **Sorgfalt**: Quellen von Fehlern und Unsicherheit im System und den Datenquellen identifizieren, dokumentieren und protokollieren, sodass Auswirkungen verstanden und Linderungsmaßnahmen entwickelt werden können.
- **Überprüfbarkeit**: Interessierten Dritten das Testen, Verstehen und Evaluieren des Systems ermöglichen, zum Beispiel durch geeignete APIs (Application Programming Interface, Schnittstellen), Information, Überprüfbarkeit ermöglichende Nutzungsbedingungen.
- **Fairness**: Sicherstellen, dass algorithmische Entscheidungen für unterschiedliche Demographien (Geschlecht, Herkunft etc.) nicht systematisch ungerecht verschiedenen Output liefern.

Diese ethischen Prinzipien sind sehr abstrakt gehalten und nicht abschließend – Aspekte wie Datenschutz und der Umgang mit Experimenten mit menschlichen Probanden fehlen zum Beispiel, auch weil die US-Rechtskultur, aus der dieser Vorschlag stammt, eine andere ist. Trotz dieser zu füllenden Lücken überzeugt der Ansatz einer prozessbezogenen Professionsethik (FAT/ML 2016).

Vergleichbare prozessbezogene professionsethische Prinzipien hat die wissenschaftliche Informatikgesellschaft Association for Computing Machinery (ACM) zur Diskussion gestellt (ACM 2017).

Anders als der prozessbezogene will der **ergebnisbezogene Ansatz einer Professionsethik auch** die erreichten Ziele eines Systems in die Bewertung einbeziehen – den „ultimate impact of an AI system" (Campolo et al. 2017: 33). Ein Beispiel dafür sind die drei allgemeinen Grundsätze im Entwurf „Ethical Alligned Design" des internationalen Berufsverbands von Ingenieuren aus den Bereichen Elektrotechnik und Informatik IEEE (Institute of Electrical and Electronics Engineers): „1. Embody the highest ideals of human rights. 2. Prioritize the maximum benefit to humanity and the natural environment. 3. Mitigate risks and negative impacts as AI/AS evolve as sociotechnical systems" (a. a. O.: 15).

Eine Konkretisierung der Idee professionsethischer Kodizes sollte auch diese Frage adressieren: Wie ist bei den Anforderungen an die Inhalte der Kodizes und an die Mittel zur Sicherung der Verbindlichkeit zwischen privatem und staatlichem Einsatz zu differenzieren? Der Staat ist unmittelbar gehalten, Gleichheit und Teilhabe zu gewährleisten, die gesetzlichen Anforderungen an Private sind niedriger. Sollten Kodizes diese Differenzierung berücksichtigen? Sollten Kodizes die potenzielle Wirkung der Systeme im Einsatzfeld beachten? Das erscheint zumindest in dem Bereich sinnvoll, wo private Betreiber öffentliche Funktionen erfüllen (z. B. Schaffung der Infrastruktur für den gesellschaftlichen Diskurs).

3. Wie erlangen die Prinzipien einer Professionsethik Verbindlichkeit und wie werden sie aktualisiert?
Häufig schlagen Experten vor, Professionsethik zu einem Teil der relevanten Ausbildung in Mathematik, Data Science, Machine Learning, Informatik und anderen relevanten Fächern zu machen. Die Herausforderungen beginnen bei der Frage, in welchen Fächern diese Inhalte verpflichtend sein sollten. Das liegt daran, dass viele Berufsgruppen an der Entwicklung von algorithmischen Systemen beteiligt sind und dass es beispielsweise für Data Scientists keine formale Ausbildung gibt. In diesem Punkt lässt sich das Vorbild der Pflichtvorlesungen zur Medizinethik im Medizinstudium nicht ohne Weiteres übertragen. Die Royal Society (2017: 12) schlägt zum Beispiel „relevant training in ethics" für „postgraduate students in machine learning" vor. Die Expertenkommission der Obama-Regierung zu Künstlicher Intelligenz empfahl hingegen einen thematisch und organisatorisch breiteren Ansatz:

> *„Schools and universities should include ethics, and related topics in security, privacy, and safety, as an integral part of curricula on AI, machine learning, computer science, and data science"* (Executive Office of the President et al. 2016: 34).

Bei der Frage der Ausbildung ist im nächsten Schritt zu konkretisieren, wer was lernen sollte. Es braucht zudem Ideen, wie nicht mehr in Ausbildung befindliche Menschen mit dieser Professionsethik vertraut gemacht werden können. Das ist auch relevant, weil viele als Data Scientists tätige Menschen aus Naturwissenschaften kommen. Ebenfalls zu konkretisieren ist, wie eine Professionsethik auf neue Fälle angewendet ihre Prinzipien entsprechend erweitert oder aktualisiert werden können. Auch hier gibt es in vielen Bereichsethiken mögliche Vorbilder: Fallbesprechung in interdisziplinären Teams in klinischen Ethikkomitees zum Beispiel. Oder die im Presserat institutionalisierte Beratung von Fällen in Ausschüssen sowie die in einem klar geregelten Prozess mögliche Aktualisierung des Kodex. Solche Beispiele erfolgreicher Professions- und Bereichsethiken müssen analysiert und wo möglich auf den Bereich algorithmische Systeme übertragen werden.

4.2 Umsetzung von Zielen in Systemen prüfen

Das voranstehende Kapitel beleuchtet Lösungsvorschläge zur Sicherstellung gesellschaftlich angemessener Ziele algorithmischer Entscheidungssysteme. Doch damit ist natürlich weder gesichert, dass die intendierten Wirkungen erzielt werden, noch, dass nicht intendierte Wirkungen ausgeschlossen werden. Gerade bei selbstlernenden Systemen gilt: Es können nicht nur Fehler oder Bugs auftreten, sondern auch unerwartete Interaktionen zwischen Bestandteilen solcher komplexen Systeme. Dazu kommt die Möglichkeit, dass durch probabilistische Entscheidungssysteme der Einzelne zum Opfer von Wahrscheinlichkeiten wird. Das gilt insbesondere für Prognosen, die nicht nur das individuelle Verhalten, sondern auch dasjenige von Freunden, Familie oder Netzwerkkontakten berücksichtigen, wie es aktuell etwa aus dem Bereich der Delinquenzprognose bekannt ist. Damit werden Überprüfung, Kontrolle oder Aufsicht über algorithmische Entscheidungssysteme notwendig (FAT/ML 2016; Future of Privacy Forum 2017; Web Foundation 2017). Sie sind die Grundlage für die Bewertung und Evaluation von sich im Einsatz befindlichen Systemen.

Daher widmet sich dieses Kapitel den Möglichkeiten der Überprüfung algorithmischer Entscheidungssysteme. Es bietet zunächst eine Übersicht über technische und institutionelle Möglichkeiten von Algorithmenanalysen (Auditings), der sich eine Darstellung von Lösungsansätzen zur Sicherung einer adäquaten Datenbasis algorithmischer Entscheidungssysteme anschließt. Die Frage, wer über welche Kapazitäten der Prüfung verfügt, ist wesentlich. Neben der hausinternen Überprüfung der Betreiber selbst ist die Etablierung zivilgesellschaftlicher Wächterinstitutionen eine Option, die Einrichtung einer öffentlich getragenen Institution für die Zulassung von und Aufsicht über algorithmische Entscheidungssysteme eine andere. Rechtliche Restriktionen, die sich beim Prüfen algorithmischer Entscheidungssysteme stellen, sowie entsprechende Lösungsoptionen beenden den engeren Bereich der Prüfung.

4.2.1 Methoden zur Umsetzungsprüfung entwickeln

Steckbrief: Algorithmenauditing

Kerngedanke: Technische und prozessuale Methoden entwickeln, um die Funktionsweise algorithmischer Entscheidungssysteme zu überprüfen

Handlungsfeld: Umsetzung prüfen

Stakeholder: Systembetreiber, Systementwickler, Wissenschaft, Nichtregierungsorganisationen, Regulierungsstellen

Durchsetzende Akteure: Systembetreiber, Wissenschaft, Staat, Regulierungsstellen

Instrumente: technische Verfahren, Standards für Datenzugang, Gesetzgebung, Prozessstandards

Status: Ideen, erste Prototypen, Vorbilder in anderen Bereichen (qualifizierte Transparenz etwa bei Wirtschaftsprüfung)

Eine wachsende Anzahl von Beiträgen widmet sich den verschiedenen Möglichkeiten des Algorithmenauditings. Es geht dabei um das Untersuchen der Funktionalität und Wirkung algorithmischer Entscheidungssysteme (Mittelstadt 2016).

Das Algorithmenauditing oder, wie es früher hieß, die Algorithmenanalyse ist eine der Grunddisziplinen der Mathematik. Mögliche Fehlerquellen einfacher algorithmischer Entscheidungssysteme umfassen insbesondere diese Phasen der Entwicklung und Einbettung (vgl. Zweig 2018: 17):

- Algorithmendesign und Implementierung
- Methodenauswahl und Operationalisierung (Datensammlung und Datenauswahl)
- Konstruktion des Entscheidungssystems
- bei selbstlernenden Systemen: Training des Systems
- Einbettung in den gesellschaftlichen Kontext
- Revalvation des Entscheidungssystems

Lösungsmöglichkeiten sehen die Autoren in der **Prüfung des Gesamtsystems (Reverse Engineering des ADM-Systems)**, das eine Transparenz seitens der Systembetreiber für alle Systemkomponenten voraussetzt (Zweig 2016).

Die Prüfung algorithmischer Systeme kann noch komplexer sein: Denn seit gut 50 Jahren bereits bergen nicht einzelne Algorithmen, sondern komplexe Softwaresysteme das Risiko, das menschliche Verständnis und seine Kontrolle zu übersteigen. Dies ist insbesondere zurückzuführen auf:

- die Vielzahl an Programmierern, die an einzelnen Teilen eines algorithmischen Systems arbeiten, sowie
- die Interaktion unterschiedlicher Systemteile, die nicht berechenbar ist (Passig 2017).

Die Systemkomplexität steigt weiter durch die Vielzahl von miteinander interagierenden Systembestandteilen, wie sie im Regelfall mit der Integration von Technologien Künstlicher Intelligenz einhergeht. Daher stellt das oben beschriebene Verfahren nur eine Lösungsmöglichkeit für Fälle dar, in denen das algorithmische System eher einfach gebaut ist: etwa die Delinquenzprognose US-amerikanischer Gerichte (Lischka und Klingel 2017) oder die Einordnung von Arbeitslosen zum Zwecke der Vermittlung adäquater Serviceangebote, wie sie derzeit in Polen vorgenommen wird (Jedrzej, Sztandar-Sztanderska und Szymielewicz 2015). Dazu müssen die Einzelkomponenten – das zu lösende Problem, Datenbasis, Modellierung sowie Algorithmus – offengelegt werden, zumindest den mit der Prüfung betrauten Akteuren.

Komplexe algorithmische Entscheidungssysteme, deren Systembestandteile dynamisch interagieren und Technologien Künstlicher Intelligenz beinhalten, kommen aktuell beispielsweise in Bereichen wie der Erstellung von Konsumentenprofilen, der Marktsegmentierung und der Personalisierung von Information in Empfehlungssystemen vor. Diese Anwendungen liegen außerhalb der Untersuchung teilhaberelevanter Entscheidungssysteme. Innerhalb des Anwendungsbereichs dieser Studie liegt allerdings die Nutzung künstlich intelligenter Systeme bei der Grenzkontrolle, wie sie in Australien zum Einsatz kommen. Doch die Methodik bleibt zentral, ihre Überprüfung erfordert andere Methoden: Ausgangspunkt sind hier Ansätze, die an klassischen Feldversuchen bzw. an der **Analyse von Input und Output** orientiert sind (Sandvig et al. 2014). Sie basieren auf der systematischen Analyse der Daten, die einem algorithmischen Entscheidungssystem zugrunde liegen, und seinen Ergebnissen.[13] Der Analysefokus verschiebt sich dabei von dem Zugriff auf Algorithmen zu dem Zugriff auf Eingaben und Ausgaben der Systeme. Da es sich zumeist um proprietäre Systeme handelt, ist der direkte Zugriff meist ebenfalls versperrt. Externe Analysen unterliegen rechtlichen Restriktionen (vgl. Kapitel 4.2.4).[14]

[13] In diesem Bereich gibt es verschiedene Testsysteme, die Software auf ihre Korrektheit überprüfen. Experten gehen davon aus, dass für die Prüfung von Algorithmen kein einzelnes Verfahren ausreichend sein wird, Tests mit Realdaten sowie Extremwerttests, Fuzzing und statische Analyse jedoch Ausgangspunkte darstellen (Dewes 2018).

[14] Eine Sonderform der Input-Output-Analyse stellt die sogenannte „eigenschaftsbasierte Prüfung von Algorithmen“ dar: Hier geht es um die Algorithmenanalyse bzw. Charakterisierung mittels künstlich generierter Daten. Sie dürfte v. a. für Systembetreiber interessant sein, andere Anwendungsfelder sind zu überprüfen (Dewes 2018).

Für die Datenerhebung solcher Analysen eignen sich prinzipiell automatisierte Verfahren wie das **Web-Scraping** und die Datenakkumulation mittels Bots. Weil diese Methoden schnell an rechtliche Grenzen stoßen, experimentieren verschiedene zivilgesellschaftliche Institutionen mit **Datenspenden**: Hier handelt es sich um eine Art der kollaborativen Erforschung der Funktionsweise großer Plattformen (Kitchin 2016; Sandvig et al. 2014). Ein Beispiel für die **kollaborative Erforschung** der Suchalgorithmen von Google ist das Datenspendeprojekt der Initiative „AlgorithmWatch“ zur Bundestagswahl 2017: Zugrunde lag die Frage, wie stark die Suchergebnisse personalisiert sind, d. h. wie stark sie algorithmisch auf eine suchende Person angepasst sind. Zu diesem Zweck wurden Nutzer über die Medienplattform „Spiegel Online“, Tageszeitungen und Social-Media-Accounts gebeten, ihre Suchergebnisse zu 16 Suchbegriffen – Spitzenpolitiker und Parteien – zu spenden. Technisches Mittel dazu war ein sogenanntes **Plugin** – eine Browsererweiterung, welche die Daten automatisch transferierte. Bei den täglich bis zu 600 Datenspendern ließen sich nur geringe Hinweise auf eine starke Personalisierung finden, allerdings eine gewisse Regionalisierung – eine Anpassung der Suchergebnisse an lokale Besonderheiten wie etwa regionale Ortsvereine (Krafft et al. 2017).

Die Autoren der Studie räumen ein, dass sie nicht repräsentativ ist. Dennoch ist sie ein innovatives und vor allem weiterentwickelbares Projekt, das eine erste Annäherung an die Funktionsweise intransparenter algorithmischer Entscheidungssysteme bietet, welche die Wahrnehmung beeinflussen. Allerdings demonstriert die Studie gleichfalls: Die alternative Datenerhebung birgt große Probleme der Repräsentativität und Geschwindigkeit, was für eine Verpflichtung von ADM-Prozessbetreibern spricht, Beforschbarkeit in bestimmten Anwendungsfeldern zwingend und standardisiert zu ermöglichen.

Ein ganz anderer Ansatz zur Prüfung komplexer Entscheidungssysteme besteht in der Entwicklung von **Algorithmen, die erklärbare Variablen und Modelle hervorbringen** (Diakopoulos 2016; Gunning 2016). Zu diesen Methoden zählen auch die bereits ausgeführten Anwendungserläuterungen (vgl. Kapitel 4.1.2). Basierend auf mathematischen Modellen geben sie Auskunft darüber, wie komplexe algorithmische Entscheidungen im Einzelfall zu ändern sind. Die Ansätze sind innovative und konstruktive Lösungen, deren Entwicklung zu beobachten ist.

Die Qualität algorithmischer Entscheidungen kann aber nicht nur durch eine Prüfung gewährleistet werden, sondern auch durch das Setzen **technischer Standards**: Ein Beispiel dafür ist das Visavergabesystem in den USA, bei dem Zugangschancen unter Bewerbern verlost werden. Das Verfahren sieht ein Diversitätsziel vor. Eine gewisse Menge an Visa soll an Menschen aus Herkunftsländern vergeben werden, die ansonsten unterrepräsentiert wären. Wer garantiert die Zufälligkeit? Das Vergabesystem lässt sich schwer überprüfen, da sowohl die Daten aufgrund von Datenschutz als auch der Algorithmus aufgrund möglicher Manipulation geheim sind. Hier sollen **prozedurale Erfordernisse** die Qualität des Entscheidungsprozesses absichern, etwa **Verschlüsselungstechnologien** und bestimmte Formen **informationswissenschaftlicher Tests** (Kroll et al. 2017).

Der Überblick über grundlegende Ansätze zum Thema Algorithmenauditing zeigt: Es ist ein weites Feld, das nicht einfach durch gesetzliche Verpflichtungen zu regeln ist, da technische Komplexität und rechtliche Hürden einer effektiven Übersicht mitunter fundamental im Weg stehen. Dennoch bleiben die Forderungen nach einer regelmäßigen und institutionalisierten Überprüfung und Zertifizierung von Algorithmen, die der automatisierten Entscheidung dienen (Citron und Pasquale 2014; Council of Europe – Committee of experts on internet intermediaries 2017: 32; Diakopoulos 2016: 26). Lösungsmöglichkeiten bestehen insbesondere in

- Forschungsprojekten, die auf Umwegen relevantes Datenmaterial sammeln, um Input-Output-Analysen durchzuführen und somit auch ein Informationsgleichgewicht gegenüber großen Systembetreibern zu fördern, sowie
- **technischen Maßnahmen**, die die Erklärbarkeit von algorithmischen Entscheidungssystemen erhöhen und ihre Fairness absichern.

Neben der Betrachtung des ADM-Systems ist natürlich auch seine Einbettung in ADM-Prozesse zu berücksichtigen, etwa Formen des Inputs, Schnittstellen zu anderen Systemen usw.

Hinzuzufügen ist natürlich, dass Algorithmenanalysen nicht zwangsläufig eine Öffentlichkeit voraussetzen müssen, die von Systembetreibern nicht erwünscht ist. Sie können etwa auch von geschlossen Gutachtergruppen durchgeführt werden, die jenseits formaler Berichtspflichten zu Vertraulichkeit oder Geheimhaltung bezüglich analysierter Systeme oder Systembestandteile verpflichtet werden, von denen sie Kenntnis genommen haben (**qualifizierte Transparenz**). Das dürfte insbesondere in denjenigen Fällen eine Lösungsmöglichkeit darstellen, wo dem Datenschutz von Individuen oder dem Schutz vor Manipulation eine hohe Bedeutung zukommt.

Um die Möglichkeiten, Restriktionen und Notwendigkeiten für unterschiedliche Fälle algorithmischer Entscheidungssysteme zu eruieren, scheint eine **Klassifikation der Systeme** sinnvoll. Diese könnte berücksichtigen:

- die Art und Komplexität von automatisierten Entscheidungssystemen
- die Anwendungsbereiche und Betreiber (öffentlich oder privat)
- mögliche Risiken für Betroffene (Tutt 2016)

Auf dieser Basis könnten gezielt Ressourcen in die Überprüfung von algorithmischen Entscheidungssystemen fließen, die ob ihrer Tragweite oder involvierter Risiken besonderer Aufmerksamkeit bedürfen bzw. wo schlichtweg unerklärliche Fehler (a. a. O.) ausgeschlossen sein müssen. Auch könnte man sich hier auf **institutionelle Lösungen** verständigen, die eine Überprüfung bei gleichzeitigem Schutz berechtigter Interessen gewähren **(qualifizierte Transparenz)** – etwa durch einen externen Gutachterkreis, eine Behörde oder unabhängige Institutionen.

4.2.2 Qualität der Datenbasis verbessern und dokumentieren

Steckbrief: Korrekturstandards und -institutionen für Datenqualität
Kerngedanke: Methoden zur Verbesserung der Aktualität, Korrektheit und Vollständigkeit von Daten entwickeln
Handlungsfeld: Umsetzung prüfen
Stakeholder: Betroffene, Systembetreiber, Wissenschaftler
Durchsetzende Akteure: Entscheider, Systembetreiber, Staat
Instrumente: Gesetz (Implementierung), Standardentwicklung
Status: Ideen

Ein hypothetisches Beispiel veranschaulicht die Bedeutung der Datenbasis in algorithmischen Systemen: Das Unternehmen X will herausfinden, ob es Zusammenhänge zwischen dem sozialen Hintergrund seiner Mitarbeiter, deren Lebensläufen und Karriereentwicklungen in der Firma gibt. Also untersucht es die internen Personaldaten. Die dürften aber einige systematische Verzerrungen enthalten: Die überwiegende Mehrheit der in den 70er, 80er und 90er Jahren eingestellten und erfolgreichen Mitarbeiter ist männlich und hat Wehrdienst geleistet. Und kein einziger der erfolgreichen Projektmanager aus diesen Zeiträumen hat eine Zertifizierung als Scrum Master! Basierend auf diesen Daten könnte man geleisteten Wehrdienst als Merkmal einbeziehen und Scrum-Zertfikate vernachlässigen – zielführend und gesellschaftlich angemessen wäre ein solches Modell aber gewiss nicht. Das Extrembeispiel zeigt: Input- und Trainingsdaten können korrekt sein, aber dennoch Verzerrungen enthalten. Sie können auch fehlerhaft sein.

Fehler können auf Ebene einzelner Datensätze festgestellt und korrigiert werden. Dazu kann ein individuelles Recht auf **Datenauskunft und -korrektur** dienen (vgl. Kapitel 4.1.1). Diese Rechte kommen nicht nur dem Einzelnen zugute, sondern der Allgemeinheit. Citron und Pascale führen am Beispiel von Kreditscoring aus:

> *„An important question is the extent to which the public should have access to the data sets and logic of predictive credit-scoring systems. We believe that each data subject should have access to all data pertaining to the data subject. Ideally, the logics of predictive scoring systems should be open to public inspection as well. There is little evidence that the inability to keep such systems secret would diminish*

> *innovation. The lenders who rely on such systems want to avoid default – that in itself is enough to incentivize the maintenance and improvement of such systems"* (Citron und Pasquale 2014: 26).

Die Autoren der „Future of Life"-Konferenz, die der Entwicklung gemeinwohlorientierter Technologien Künstlicher Intelligenz gewidmet ist, formulieren kurz und knapp:

> *„People should have the right to access, manage and control the data they generate, given AI systems' power to analyze and utilize that data"* (Future of Life Institute 2017).

Sie gehen damit noch einen Schritt weiter und fordern ebenfalls die Möglichkeit, persönliche Daten zu korrigieren – was im Bundesdatenschutzgesetzt (BDSG) entsprechen geregelt ist („§ 34 BDSG – Einzelnorm" o. J.).

Die Idee ist simpel: Wenn jeder die Möglichkeit hat, die ihn betreffenden Daten einzusehen und ggf. zu korrigieren, ist für die Allgemeinheit gesorgt. Sie entspricht der Idee sogenannter **Data Methods Solutions** – Methoden der Datenerhebung und Auswertung, welche die Aktualität, Korrektheit und Vollständigkeit von Daten absichern. Denn diese liegen der Entwicklung algorithmischer Analyse- und Entscheidungsverfahren zugrunde. Die Kontrolle von Individuen über sie betreffende Daten bietet eine zuverlässige Option, die Qualität eines Datenkorpus in der Gesamtheit zu sichern und damit die Zuverlässigkeit algorithmischer Entscheidungen zu erhöhen (Future of Privacy Forum 2017). Allerdings ist es nicht sinnvoll, die Verantwortung hier allein bei dem Betroffenen zu verorten. Komplexe, aktuelle, große Datensätze kann der Einzelne kaum auf Aktualität, Korrektheit und Relevanz überprüfen. Hier sind Institutionen nötig.

Die Europäische Datenschutz-Grundverordnung (DSGVO-EU) enthält hierzu ebenfalls Regelungen, insbesondere in Kapitel 3, Art. 13 bis 16 im Kontext personenbezogener Daten (Europäisches Parlament und Rat der Europäischen Union 2016). Diese sehen etwa vor, dass Verantwortliche der Datenerhebung und -verarbeitung Betroffenen bei Bedarf eine Kopie personenbezogener Daten zur Verfügung stellen (Art. 15 DSGVO-EU). Außerdem hat jeder Betroffene das Recht, eine Berichtigung der Daten vornehmen zu lassen (Art. 16 DSGVO-EU).

Offene Fragen bestehen hier vor allem hinsichtlich der Durchsetzung: Um Rechte durchzusetzen, braucht es Institutionen. Dazu empfehlen Experten und Branchenverbände die Einrichtung sogenannter **Datenombudsleute** oder **Datenschutzbevollmächtigter** (Otto 2017: 27).

Steckbrief: Qualitätssiegel für Herkunft und Güte von Daten

Kerngedanke: Methoden zur Absicherung der Aktualität, Korrektheit und Vollständigkeit von Daten entwickeln

Handlungsfeld: Umsetzung prüfen

Stakeholder: Betroffene, Systembetreiber, Wissenschaftler

Durchsetzende Akteure: Entscheider, Systembetreiber, Staat

Instrumente: Gesetz (Implementierung), Standardentwicklung

Status: Ideen, erste Prototypen

Um systematischen Verzerrungen in Daten (z. B. Über- oder Unterrepräsentation bestimmter Gruppen) zu begegnen, die über einen Einzelfall hinausgehen, braucht es andere Instrumente als Korrektur- und Beschwerdemöglichkeiten. Damit müssen Regelungen zur Sicherung der Datenbasis gefunden werden, die kollektiv die Aktualität, Korrektheit und Vollständigkeit von Daten sichern und der Tatsache Rechnung tragen, dass diese global gehandelt werden. Wenn jemand nachvollziehen möchte, warum der Output eines algorithmischen Systems bestimmte Verzerrungen aufweist, muss etwas über die **Art und Herkunft der Trainingsdaten** bekannt sein. Eine hilfreiche Analogie dafür ist die **Dokumentation von Lieferketten** in der Lebensmittel- oder Bekleidungsindustrie. So wie es wichtig ist, die Herkunft von verarbeiteten Hühnereiern bis zu einzelnen Höfen nachvollziehen zu können, müssen auch die Datensätze auffindbar sein, die zum Beispiel bei der Entwicklung eines Modells zur Gesichtserkennung im Einsatz waren. Dafür braucht es Standards, die Auskunft geben über

die Herkunft und Beschaffenheit der Trainingsdatenbank und der Trainingsdatensätze (ACM 2017; FAT/ML 2016), zum Beispiel hinsichtlich der Repräsentativität, Aktualität usw. Einen ersten Vorschlag hat ein Forscherteam unter dem Schlagwort „Datasheets for Datasets“ entwickelt und prototypisch auf einige reale Trainingsdatensätze angewendet (Gebru et al. 2018).

Für die Datenerhebung bedeutet dies: Wer Trainingsdaten erhoben hat, sollte diese mit standardisierten Informationen über Beschaffenheit, absehbare Verzerrungen oder Einschränkungen bei der Verwendung kennzeichnen. Verbindliche Mindestanforderungen an die Auszeichnung bestimmter Datensätze können ein Teil der Lösung sein. Eine solche Auszeichnung dient der Erklärbarkeit und Nachvollziehbarkeit der Entscheidungen algorithmischer Systeme:

> *„Develop standards to track the provenance, development, and use of training datasets throughout their life cycle. This is necessary to better understand and monitor issues of bias and representational skews. In addition to developing better records for how a training dataset was created and maintained, social scientists and measurement researchers within the AI bias research field should continue to examine existing training datasets, and work to understand potential blind spots and biases that may already be at work“* (Campolo et al. 2017: 2).

Ein weiteres Problem sind Verzerrungsmöglichkeiten beim Datentransfer. An den Schnittstellen eines ADM-Systems zu Auskunftsdatenbanken können Verzerrungen ebenso auftreten wie durch die Software selbst. Beispiele dafür finden sich in dem äußerst sensiblen Bereich der DNA-Analyse bei der Strafverfolgung (Kirchner 2017). Hier bietet der Vergleich verschiedener Systeme einen Ausgangspunkt, den Problemen beizukommen.

Zu erörtern ist, ob eine Regulierung dahingehend sinnvoll ist, dass das „Training“ von Algorithmen an eine Anwendung im gleichen oder zumindest in einem vergleichbaren Kontext gebunden wird. Ein Beispiel wäre, dass in der Medizin eine Diagnosesoftware, die mithilfe von Trainingsdaten entwickelt wurde, nur bei Populationen eingesetzt werden sollte, die mit derjenigen übereinstimmen, die die Trainingsdaten geliefert hat. Zentral ist die Übereinstimmung von Aspekten, die für die Zielstellung der Software relevant sind. Sind diese noch nicht bekannt, sollte die Software nicht in einem neuen, noch nicht hinreichend ergründeten Kontext eingesetzt werden (Otto 2018).

4.2.3 Überprüfbarkeit algorithmischer Systeme gesetzlich ermöglichen und sichern

Steckbrief: Rechtliche Hürden für Überprüfbarkeit senken

Kerngedanke: Rechtliche Restriktionen für die Überprüfung algorithmischer Entscheidungssysteme müssen für Forschungszwecke abgeschafft bzw. limitiert werden (z. B. nach Katalogkriterien)

Handlungsfeld: Umsetzung prüfen

Stakeholder: Entscheider

Durchsetzender Akteur: Staat

Instrument: Rechtsetzung

Status: Ideen

Privatrechtliche, urheberrechtliche und sicherheitsrechtliche Normen sollten so angepasst werden, dass die Überprüfung von algorithmischen Entscheidungssystemen durch externe Akteure möglich ist. Aktuell begeben sich zum Beispiel Sicherheitsforscher in den Bereich des Strafrechts, wenn sie Funktionen und Sicherheit automatisierter Systeme überprüfen. Langfristig würde ein solcher Zustand gerade in teilhaberelevanten Bereichen dazu führen, dass Missbrauch und Fehler schwer nachvollziehbar sind.

Ausgangspunkt für alle rechtlichen Restriktionen ist die Tatsache, dass Algorithmen häufig einer Geheimhaltung unterliegen. Trainings- und Anwendungsdaten unterliegen im Regelfall einer privaten Verfügungsgewalt. Das hat unter Umständen gravierende gesellschaftliche Auswirkungen:

In der Folge des Brexit-Referendums und der Wahl Donald Trumps zum Präsidenten der USA gab es eine Debatte unter Experten, inwieweit automatisierte Technologien der Wahrnehmungsbeeinflussung zu den Entscheidungen beigetragen haben. Einerseits stand die Vermutung im Raum, dass soziale Netzwerke durch Algorithmen sogenannte **Echokammern** verstärken, in denen Individuen vor allem mit solchen Inhalten in Kontakt kommen, die ihren persönlichen Präferenzen entsprechen. Dies führt über Zeit zu einer Polarisierung der Gesellschaft; eine Entwicklung für die es bereits es Anhaltpunkte gibt. Gleichfalls zur Debatte stand der Einfluss sogenannter **Fake News** (Falschnachrichten mit manipulativer Intention), die mithilfe automatisierter Verfahren (Social Bots) und Technologien äußerst gezielter Werbemaßnahmen (Microtargeting) eine Veränderung des gesellschaftlichen Meinungsklimas herbeiführen konnten.

Leider konnte die Wirkung von Algorithmen und automatisierten Technologien auf die Onlinedebatten bislang nicht umfassend untersucht werden, da die betreffenden sozialen Netzwerke keine unabhängigen Studien zulassen (Lischka und Stöcker 2017). Die einzige Studie zu vermeintlichen Echokammern, bei der entsprechende Nutzerdaten direkt ausgewertet werden konnten, wurde von Mitarbeitern von Facebook selbst durchgeführt. Sie sollte belegen, dass die Präferenzen der Nutzerinnen und Nutzer einen größeren Einfluss auf die Nachrichtenselektion (Newsfeed) haben als Algorithmen. Der Einfluss von Algorithmen auf die Nachrichtenselektion und das Ranking der Nachrichten war einerseits nachweisbar, blieb in seiner Bedeutung allerdings unklar. Die Studie barg elementare Probleme hinsichtlich ihrer Repräsentativität: Sie untersuchte ausschließlich Nutzer, die Facebook einerseits viel nutzen und sich dort andererseits politisch identifiziert haben (2 Prozent der Nutzer). Die Auswahl der Nutzer interagierte mit der Fragestellung. Die Studie ergab eine breite Debatte über ihre Methodik, allerdings keine zufriedenstellenden Forschungsergebnisse (Sandvig 2015).

Das Beispiel zeigt: Der rechtliche Status quo im Bereich Algorithmen und Daten kann unter Umständen gesamtgesellschaftliche Schäden hervorrufen. Abhilfe böte hier eine Einschränkung von Geschäftsgeheimnissen zum Zwecke der Erforschung entsprechender Systeme im Einzelfall und unter Berücksichtigung berechtigter Interessen (Calo 2017) bzw. der Erlass entsprechender Transparenzvorschriften (Tutt 2016).

Neben der Verpflichtung auf die Bereitstellung relevanter Informationen kann auch die **externe Beforschbarkeit** gestärkt werden. Diese beruht maßgeblich auf automatisierten Datenerhebungsmethoden wie dem Web (automatisierte Sammlung öffentlich verfügbarer Informationen) oder der Einrichtung multipler (automatisierter) Nutzeraccounts als Grundlage für **Input-Output-Analysen** (Kapitel 4.2.1). Diese Methoden verletzen im Regelfall die Allgemeinen Geschäftsbedingungen der Internetkonzerne. Durch **IT-Sicherheitsgesetze** wie den „Computer Fraud and Abuse Act“ (1986) galten sie bis zum Januar 2018 als nicht autorisierter Zugriff in vernetzte Systeme und waren Gegenstand des Strafrechts (Calo 2017; Stone et al. 2016): Aktuell sind in den USA mehrere Gerichtsverfahren anhängig, die darauf abzielen, juristische Unklarheiten zu beseitigen und eine entsprechende Forschung zu schützen. Denn unabhängige Input-Out-Analysen sind bislang die wichtigste Methode, Diskriminierung in algorithmischen Prozessen aufzudecken (ACLU 2017; Goodman 2015; Williams 2017). Am 10. Januar 2018 erging ein Urteil des Bundesberufungsgerichts für den 9. District der USA, das die bisherige Rechtsprechung kippte: Die Verletzung Allgemeiner Geschäftsbedingungen von Webseiten stelle keine Straftat dar (Williams 2018). Die Gesetzgebung in der Europäischen Union und in Deutschland wäre hierzu zu prüfen. Der aktuell dem Bundestag zur Beratung vorliegende Entwurf zur Einführung des neuen Straftatbestandes **digitaler Hausfriedensbruch** würde beispielsweise analog dem „Computer Fraud and Abuse Act“ entsprechende Forschung kriminalisieren (Buermeyer 2016; juris 2018).

Ebenfalls betroffen ist das **Urheberrecht** (a. a. O.), insbesondere die sogenannte **Anti-Circumvention Provision** des „Digital Millennium Copyright Act“. Potenzielle Rechtsverletzungen sind der Verletzung von IT-Sicherheitsgesetzen ähnlich, auch hier geht es um die automatisierte Datenerhebung mittels automatisierter Nutzeraccounts bzw. Bots: Beispielsweise kann ein getarnter Account („Mensch mit afroamerikanischer Herkunft“) dazu eingesetzt werden, die adäquate Funktionsweise von Systemen zu testen, die Gesichtserkennung zum Gegenstand haben oder als Methode zu einem anderen Zweck benutzen. Zahlreiche Beispiele belegen, dass Gesichtserkennungssoftware solche Menschen mit afroamerikanischer Herkunft häufig nicht erkennt (AI Now Institute 2017) oder falsch klassifiziert – etwa als Gorilla (Doctorow 2018). Eine Prüfung erscheint insofern sehr

sinnvoll. Das Verfahren, d. h. der Einsatz von automatisierten Accounts, erfordert allerdings eine Validierung des Accounts seitens der Systembetreiber zum Zwecke der Identitätsüberprüfung. Dessen Fälschung zum Zwecke der Forschung kann die oben genannten Normen in der Hinsicht verletzen, als dass diese sich auch auf Technologien erstrecken, die Zugriffsbeschränkungen auf kopiergeschützte Werke umgehen – selbst wenn dabei keine Urheberrechtsverletzungen begangen werden. Hier ist die lokale Gesetzgebung ebenfalls auf den Prüfstand zu stellen

Grund zur Sorge bietet außerdem die aktuelle Überarbeitung der „Datenbanken-Richtlinie der EU“ im Kontext der EU-Urheberrechtsreform, welche diversen Stakeholdern Rechtssicherheit im Kontext offener Daten verwehrt. Dies ist für das Algorithmenauditing insofern relevant, als Datenbanken mit offenen Daten eine Grundlage für vergleichende Algorithmentests und Modellbildung darstellen könnten (vgl. Kapitel 4.3.1 und 4.3.2).

Die Überprüfbarkeit algorithmischer Systeme kann durch **Dokumentationspflichten der Betreiber** steigen. Gemeint ist die Dokumentation insbesondere zugrunde liegender Modelle, Algorithmen und Daten (ACM 2017; Shneiderman 2016), die bei Bedarf der Öffentlichkeit oder Regulierern zugängig sein muss.

4.2.4 Widerspruchsmöglichkeiten bei algorithmischen Prozessen institutionalisieren

Steckbrief: Individuelle Widerspruchverfahren

Kerngedanke: Individuelle und überindividuelle Widerspruchsmöglichkeiten gegen ADM-Verfahren, wie zum Beispiel Verbandsklagerechte für Verbraucherverbände, stärken

Handlungsfeld: Umsetzung prüfen

Stakeholder: ADM-Betreiber

Durchsetzender Akteur: Staat

Instrumente: Rechtsetzung, Institutionen

Status: Ideen

Algorithmische Entscheidungssysteme stellen rechtsstaatliche Verfahren vor große Herausforderungen, beispielsweise im Bereich der **Anfechtbarkeit von Entscheidungen**. Es ist absehbar, dass die Entscheidungsprozesse mancher ADM-Systeme auditfest dokumentiert werden müssen, um Entscheidungsforensik zu ermöglichen

Die EU-Datenschutz-Grundverordnung (DSGVO-EU) sieht im Kontext algorithmischer Entscheidungsverfahren in Kapitel 3, Art. 21 bis 22 sowohl **individuelle Widerspruchsrechte** vor als auch **das Recht auf menschliche Intervention** seitens des Verantwortlichen und auf **Darlegung des eigenen Standpunktes** (Europäisches Parlament und Rat der Europäischen Union 2016). Diese Rechte gelten Wissenschaftlern als zentral (Citron 2008). Eine umfassende Bewertung der DSGVO-EU mit Blick auf algorithmische Prozesse liefern Dreyer und Schulz (2018).

Steckbrief: Verbandsklagerechte bei algorithmischen Prozessen

Kerngedanke: Individuelle und überindividuelle Widerspruchsmöglichkeiten gegen ADM-Verfahren, wie zum Beispiel Verbandsklagerechte für Verbraucherverbände, stärken

Handlungsfeld: Umsetzung prüfen

Stakeholder: ADM-Betreiber

Durchsetzender Akteur: Staat

Instrumente: Rechtsetzung, Institutionen

Status: Ideen

Generell gilt es, Widerspruchsrechte nicht mit Blick auf eine individuelle Durchsetzung zu institutionalisieren. Auch die Möglichkeit der Einrichtung von **Verbandsklagerechten** sollte erörtert werden. Verbandsklagerechte für Verbraucherschützer erweisen sich im Datenschutz als starkes Mittel zur Durchsetzung kollektiver Interessen. Insbesondere wenn gruppenbezogene Werte verletzt sind (etwa durch Diskriminierung), ist dies ein überzeugendes Werkzeug. Was Spindler in einer Stellungnahme zur Erweiterung der Verbandsklagebefugnisse auf alle datenschutzrechtlichen Normen bzw. Verstöße gegenüber Verbrauchern in Deutschland ausführt, lässt sich auch auf algorithmische Systeme übertragen: Zivilrechtliche Klagemöglichkeiten für Verbände können die öffentlich-rechtliche Überwachung ergänzen:

> *„Die schon jetzt im BDSG gegebenen Möglichkeiten einer zivilrechtlichen Klage auf Schadensersatz, Unterlassung etc. für den einzelnen Betroffenen reichen in der Praxis kaum aus, um auf zivilrechtlicher Ebene die Einhaltung des Datenschutzrechts durchzusetzen, da der Einzelne kaum Anreize und auch nicht die nötigen Informationen hat, seine Rechte selbst durchzusetzen – von Problemen der Berechnung des Schadensersatzes etc. abgesehen. Daher ist es durchaus zutreffend, wenn darauf hingewiesen wird, dass der Einzelne sich oftmals nur durch Beschwerden bei den Datenschutzaufsichtsbehörden zu helfen weiß – daraus folgt aber nicht, dass deswegen nur und allein diese tätig werden sollten, vielmehr, dass durch Verbandsklagebefugnisse eine Entlastung der Aufsichtsbehörden eintreten kann“* (Spindler 2015: 2).

Eine Stärkung des Datenschutzes durch Verbandsklagerechte diagnostiziert Roßnagel (2017: 38) aufgrund einer ähnlichen Wirkungslogik auch in einer Analyse der DSGVO-EU: „Das Verbandsklagerecht wird als wesentliche Stärkung des Datenschutzes angesehen, wird aber für die Aufsichtsbehörden eine deutliche Mehrarbeit bedeuten.“

Mit Blick auf algorithmische Systeme ist die Rolle von **Antidiskriminierungsverbänden** und der **Antidiskriminierungsstelle des Bundes (ADS)** zu prüfen und ggf. zu erweitern. Beide haben bislang keine Verbandsklagerechte. Dieses Instrument kann sinnvoll sein, wenn Diskriminierungsmuster auf kollektiver stärker als auf individueller Ebene zutage treten und wirken. Das unabhängige Gremium zur Evaluation des Allgemeinen Gleichbehandlungsgesetzes (AGG) schlug 2016 beispielsweise ein **altruistisches Klagerecht** sowie ein **kleines Verbandsklagerecht** für die ADS vor:

> *„In Fällen öffentlichen Interesses, wenn also keine von konkreten Benachteiligungen Betroffene bekannt sind, sollte der ADS ein altruistisches Klagerecht eingeräumt werden, mit dem die Feststellung eines Verstoßes gegen das AGG vorgenommen werden kann. Das Feststellungsurteil soll in späteren individuellen Klagen Indizwirkung haben“* (Berghahn et al. 2016: 191).

> *„Ebenfalls denkbar wäre ein kleines Verbandsklagerecht der ADS, mit dem Verstöße gegen das AGG – im Einverständnis mit unmittelbar Betroffenen – geltend macht werden, ohne individuelle Rechtsansprüche durchzusetzen. Ziel wäre hier vielmehr die Feststellung der grundsätzlichen Rechtswidrigkeit eines Verhaltens bzw. einer Regelung“* (a. a. O.: 195).

4.2.5 Öffentliche Aufsicht über algorithmische Systeme entwickeln

Die vorangegangenen Kapitel konzentrieren sich auf die Erfordernisse und Lösungsansätze im Bereich der Prüfung algorithmischer Entscheidungssysteme. Ausgeklammert wird dabei zumeist die Frage, wer für eine Prüfung bzw. ihre Teilbereiche zuständig ist oder zuständig sein sollte.

Steckbrief: Bedarf und Instrumente für verschiedene Aufsichtsansätze konkretisieren

Kerngedanke: Nach Aufsichtsart (vorab, fortwährend, nachträglich), nach sektorspezifischen und nach übergreifenden Ansätzen konkretisieren

Handlungsfeld: Umsetzung prüfen

Stakeholder: Entscheider, Systembetreiber

Durchsetzende Akteure: Staat, Entscheider

Instrumente: Institutionen, Zulassungsverfahren, Siegel, Zertifikate

Status: Ideen

Verschiedene vorgestellte Lösungsansätze implizierten die Notwendigkeit der Einrichtung staatlicher Aufsichtsstellen oder zumindest die Definition und Übertragung von Zuständigkeiten an bestehende Institutionen. Im Kern geht es dabei einerseits darum, diejenigen algorithmischen Entscheidungssysteme zu identifizieren, die aus gesamtgesellschaftlicher Perspektive eine Prüfung erfordern. Andererseits müssen dann wirksame Verfahren entwickelt werden, um eine solche Prüfung zu ermöglichen. Ob Zulassungsbehörde, Algorithmen-TÜV oder Ratingagentur (Otto 2017) – Übersicht und Kontrolloptionen müssen in Kooperation mit Systembetreibern entwickelt werden. Bisherige Vorschläge sind wenig konkret. Für eine Konkretisierung bedarf es einer systematischen Analyse, die den Status quo erfasst und diese Fragen beantwortet:

1. Spricht etwas substanziell dagegen, die folgenden Analysepunkte auf in Deutschland im Einsatz befindliche algorithmische Prozesse zu beschränken (unser Vorschlag zur praktikablen Einschränkung)?
2. In welchen Sektoren sind algorithmische Systeme heute oder perspektivisch im Einsatz?
3. In welchen dieser Sektoren ist eine Aufsicht über algorithmische Systeme sinnvoll?
4. Welche Aufsichtsinstitutionen gibt es in diesen Sektoren?
5. Wo sind aus dem Abgleich der voranstehenden Fragen 2, 3 und 4 Aufsichtslücken erkennbar?
6. Welche Formen der Aufsicht (vorab/fortwährend/nachträglich) sind in den Einsatzbereichen derzeit institutionalisiert?
7. Welche Formen der Aufsicht sind bei algorithmischen Systemen sinnvoll?
8. Wo sind aus dem Abgleich der voranstehenden Fragen 6 und 7 Aufsichtslücken erkennbar?

Nur auf einer solchen empirischen Basis (ggf. zu fokussieren auf bestimmte besonders teilhaberelevante Bereiche) über die Funktionsweise bestehender Aufsichtsinstitutionen in alle relevanten Sektoren (von der Flugunfalluntersuchung über die Medikamentenzulassung bis zur Finanzaufsicht und Kfz-Zulassung) kann eine zielführende Diskussion über den Bedarf algorithmischer Entscheidungssysteme und entsprechender Lösungsstrategien geführt werden. Braucht es sektorspezifische Lösungen oder sollten Können und Kompetenzen zentral gebündelt werden (vgl. dazu Kapitel 4.4.2)? Wo sind Zulassungsverfahren sinnvoll, wo genügt eine nachträgliche Prüfung des Einsatzes?

Algorithmenfolgeabschätzungen bzw. **-verträglichkeitsprüfungen** (vgl. Kapitel 4.1.3) können eine Grundlage für die erste Einschätzung algorithmischer Systeme sein. Informationen zu den Zielen und Methoden eines algorithmischen Entscheidungssystems, der Qualität des Dateninputs sowie erwarteter Ergebnisse können die Klassifikation von Entscheidungssystemen ermöglichen, auf Basis derer verfügbare Ressourcen/Analysekapazitäten mit der Überprüfung relevanter Fälle verknüpft werden können. Die **Klassifikation der Entscheidungssysteme** sollte berücksichtigen:

- Art und Komplexität algorithmischer Systeme

- Anwendungsbereich und Betreiber (öffentlich oder privat)
- mögliche Risiken für Betroffene (Tutt 2016)

Die Durchführung relevanter Analysen durch eine staatliche Aufsichtsbehörde stellt gleichzeitig eine Option dar, qualifizierte Transparenz für die Überprüfung sensibler Bereiche zu sichern. Sie sollte nicht nur algorithmische Systeme selbst umfassen, sondern auch auf relevante Trainingsdaten sowie die Entscheidungsstrukturen, in welche die Systeme eingebettet sind.

4.2.6 Zivilgesellschaftliches Engagement fördern

Steckbrief: Zivilgesellschaftliche Wächterorganisationen etablieren

Kerngedanke: Zivilgesellschaftliche Wächter decken neuartige Problembereich automatisierter Entscheidungen auf, erproben Lösungsmöglichkeiten

Handlungsfelder: Umsetzung prüfen, Optimierungsziele auf Angemessenheit prüfen

Stakeholder: Entscheider, zivilgesellschaftliche Institutionen

Durchsetzende Akteure: Staat, Nichtregierungsorganisationen

Instrumente: Förderung, Rechtsetzung

Status: erste Akteure aktiv (z. B. AlgorithmWatch[15])

Nicht nur der Staat kann eine Wächterfunktion ausüben. Zivilgesellschaftliche Akteure haben sich in vielen Bereichen Renommee bei der Aufdeckung neuartiger Fehler und Probleme von Softwaresystemen erarbeitet. Das gilt auch für algorithmische Entscheidungssysteme. Oft waren sie es, die Probleme der Öffentlichkeit und damit auch der Politik bekannt machten.

Der sogenannte **Heartbleed-Exploit** in der Software OpenSSL zeigte 2014 anschaulich, dass Transparenz allein nicht zu Öffentlichkeit führt (Kroll et al. 2017: 10). Die gravierende Sicherheitslücke in OpenSSL, einer Standardsoftware zur Verschlüsselung von Datentransport im Internet, blieb über Jahre nicht öffentlich, obgleich der Quellcode allgemein zugänglich war. Nach Bekanntwerden der Lücke machte die OpenSSL Foundation darauf aufmerksam, dass die Pflege und Entwicklung der Software im Wesentlichen Aufgabe eines einzigen hauptamtlichen Mitarbeiters war. Damit Transparenz zu Öffentlichkeit führt, braucht es finanzielle und personelle Ressourcen. Eine Lehre aus dem Heartbleed-Exploit war die Etablierung der **Core Infrastructure Initiative (CII)**: Die Linux Foundation und große Internetunternehmen finanzieren Entwicklerstellen und Audits für die Software-Infrastruktur des Netzes (Diedrich 2014).

Die Wächterfunktion betrifft nicht nur die Technik: Bei der Analyse von Prozessen algorithmischer Entscheidungsfindung sind erschwerende Faktoren wie die Geheimhaltung verwendeter Verfahren und Trainingsdaten gängig. Wie der australische Centrelink-Fall (Rohde 2017) zeigt, können zivilgesellschaftliche Wächter auch im Vorfeld technischer Analyse eine entscheidende Funktion ausüben. In Australien wurde eine Software entwickelt, die Sozialbezüge dahingehend überprüft, wer zu Unrecht Arbeitslosengeld oder Sozialhilfe bezogen hat. Ein Computerprogramm sollte die Fälle prüfen und im Verdachtsfall automatisch beheben. Es verschickte automatisch Mahnungen, lag bei einer ersten Analyse allerdings mindestens 20.000 Mal daneben. Betroffene mussten Monate kämpfen, um Fehler zu korrigieren. Hier wurde die Debatte über die Wirkweise des Systems initiiert, indem Betroffenenorganisationen die Fälle einzelner Fehlentscheidungen aggregierten und den Medien bekannt machten.

[15] Transparenzhinweis: Die Bertelsmann Stiftung (2017) fördert die Organisation AlgorithmWatch („Förderung – Mehr Transparenz und zivilgesellschaftliche Kontrolle von Algorithmen“).

Auch die Debatte über das prominente COMPAS-System zur Risikoeinschätzung von Straftätern begann erst, nachdem das gemeinnützige US-Recherchebüro Propublica mit großem Aufwand Daten recherchiert, aufbereitet und ausgewertet hatte (Angwin et al. 2016).

Die Beispiele zeigen: Zivilgesellschaftliche Wächterfunktionen benötigen Ressourcen nicht nur für die technische Analyse von Systemen wie bei OpenSSL, sondern auch für die Recherche potenzieller Fälle, für das Sammeln und Einklagen von Daten, die Recherche tatsächlicher Einsatzpraktiken, aber auch für die Entwicklung von Standards für den Datenaustausch, für Testdatenbanken und vieles mehr.

Weitere Handlungsbereiche werden mit Blick auf Vorbilder aus dem Bereich der Umweltbewegung sichtbar: Hier fungieren Nichtregierungsorganisationen zum Teil als Beschwerdeinstanzen mit niedrigen Zugangshürden. In der Vergangenheit illustrierten sie allgemeine Probleme gleichfalls durch die Aggregation und Auswertung einzelner Fälle, machten sie öffentlich und stießen damit eine gesellschaftliche Debatte an. Die Etablierung zivilgesellschaftlicher Wächter im Bereich algorithmischer Entscheidungssysteme schließt damit an Forderungen nach Verantwortlichkeit gegenüber der Öffentlichkeit:

> *„After releasing an AI system, companies should continue to monitor its use across different contexts and communities. The methods and outcomes of monitoring should be defined through open, academically rigorous processes, and should be accountable to the public. Particularly in high stakes decision-making contexts, the views and experiences of traditionally marginalized communities should be prioritized“* (Campolo et al. 2017: 1).

Die Erfahrungen mit Fällen wie COMPAS und Centrelink zeigen, dass zivilgesellschaftliche Wächter die notwendigen Wächterfunktionen ausüben können. Um dies dauerhaft zu tun, müssen geeignete Organisationen mit relevanten Zielen, nachgewiesener Kompetenz, transparenter und geeigneter Methodik und entsprechendem Know-how identifiziert und gefördert werden.

4.3 Diversität schaffen

Vielfalt algorithmischer Systeme umfasst zwei Ebenen: Einerseits die Vielfalt der Umsetzung und damit der Systeme einem Einsatzfeld (Kapitel 4.3.1). Andererseits die Vielfalt der Ziele und damit auch der Betreiber aus dem öffentlichen, privatwirtschaftlichen und zivilgesellschaftlichen Sektor (Kapitel 4.3.2 und 4.3.3).

Vielfalt auf beiden Ebenen ist aus folgenden Gründen wichtig:

- **Schadenspotenzial eines Systems begrenzen**: Systemvielfalt auf beiden Ebenen (Ziele und Umsetzung) eröffnet Betroffenen Ausweichmöglichkeiten. Die Fehler eines Systems treffen Menschen weniger hart, wenn anders wirkende Systeme existieren.
- **Innovation:** Wenn unterschiedliche Ansätze konkurrieren, begünstigt das Neuerungen.
- **Fehlerkorrektur**: Abweichungen bei Zielerreichung und Wirkung zwischen unterschiedlichen Methoden bringen neue Erkenntnisse und fördern den Wettbewerb.
- **gesellschaftliche Dynamik:** Die Vielfalt von algorithmischen Systeme mit unterschiedlichen Zielen, Betreibern und Operationalisierungen begünstigt, dass gesellschaftliche Entwicklungen sich schneller zumindest in einigen der Systeme niederschlagen.
- **Umfassende Gemeinwohlorientierung:** Nicht jedes gesellschaftlich sinnvolle Ziel ist mit gewinnorientiertem Geschäftsmodell sinnvoll umzusetzen. Deshalb ist die Vielfalt von Betreibern und Optimierungszielen wichtig: Öffentliche, privatwirtschaftliche und zivilgesellschaftliche Akteure haben unterschiedlichen Ziele, Zielgruppen und Funktionsweisen, sie folgen unterschiedlichen Prinzipien und bearbeiten unterschiedliche Probleme. Die Vielfalt all dieser Akteure ist Grundlage für eine umfassende Gemeinwohlorientierung beim Einsatz algorithmischer Systeme.

4.3.1 Vielfalt durch zugängliche Trainingsdatensätze stärken

Steckbrief: Trainingsdatensätze durch öffentliche Förderung und Regulierung zugänglicher machen

Kerngedanke: Eine Voraussetzung für die Vielfalt algorithmischer Entscheidungssystemen ist die Zugänglichkeit relevanter Trainingsdatensätze für unterschiedliche Anbieter

Handlungsfeld: Diversität

Stakeholder: ADM-Entwickler, ADM-Betreiber

Durchsetzende Akteure: staatliche Nichtregierungsorganisationen (Forschungsförderung)

Instrumente: rechtliche Förderung (Forschung), Aufbau von Institutionen

Status: Idee

Für die enormen Fortschritte bei der Bilderkennung durch algorithmische Systeme in den vergangenen Jahren ist neben neuer Hardware auch die breite Verfügbarkeit von Trainingsdaten verantwortlich. Im Jahr 2009 veröffentlichte ein Team um die Informatikerin Fei-Fei Li den Imagenet-Datensatz – eine Datenbank von damals 3,2 Millionen verschlagworteten Fotos. Menschen hatten über Amazons Crowdsourcingplattform „Mechanical Turk" klassifiziert, was auf den Fotos zu sehen ist. Dieser Datensatz eignet sich hervorragend als Trainingsmaterial, um schwache Künstliche-Intelligenz-Muster bzw. -Zusammenhänge suchen und in einem Modell abbilden zu lassen, mit dem der Inhalt neuer Fotos automatisch bestimmt werden kann (Deng et al. 2009). Die Schöpferin des Wettbewerbs Li spricht von einem Paradigmenwechsel, den der Erfolg des Wettbewerbs ausgelöst hat – hin zum teilautomatisierten Lernen anhand von Trainingsdaten:

> *„The paradigm shift of the ImageNet thinking is that while a lot of people are paying attention to models, let's pay attention to data. Data will redefine how we think about models"* (Gershgorn 2017: 1).

Inzwischen umfasst der Datensatz 13 Millionen Fotos, die Erkennungsrate der besten Software im jährlichen Imagenet-Wettbewerb ist von 71,8 Prozent im 2010 auf 97,3 Prozent im Jahr 2017 gestiegen (a. a. O.). Das belegt anschaulich die Bedeutung von Trainingsdaten. Die 2010 erstmals am Imagenet-Datensatz erprobten Verfahren waren nicht völlig neu, Ansätze wie künstliche neuronale Netze waren da schon lange bekannt. Neu war ein Datensatz mit diesem Umfang, dieser Qualität und derart leichter Zugänglichkeit.

Zwei wichtige Einschränkungen: Die Ergebnisqualität ist nicht perfekt, wie die Performance von 97,3 Prozent suggerieren könnte. Je nach Hautfarbe zum Beispiel gibt es erhebliche Abweichungen nach unten. Und die Leistungsverbesserung ist auch von vielen anderen Faktoren begünstigt worden: bessere Kamera zum Beispiel oder andere Beleuchtung.

Im Fall der automatischen Gesichtserkennung zeigt sich der Wert großer, proprietärer Stichproben besonders deutlich. Zwar gibt es öffentlich verfügbare Stichproben, die vor allem als Teststichproben von Bedeutung sind (z. B. eine Sammlung von 13.000 Aufnahmen von Prominenten „in freier Wildbahn" – „Labeled Faces in the Wild"). Die mit Abstand besten Erkennungsraten werden auf diesen Stichproben jedoch von Konzernen wie Facebook und Google erzielt, die die Möglichkeit besitzen, viele Millionen Bilder ihrer Nutzer für das Training ihrer Gesichtserkennungssysteme zu verwenden. Eine Veröffentlichung dieser konzerneigenen Stichproben ist schon aus Datenschutzgründen nicht denkbar.

Ein weiteres Beispiel für wertvolle Trainingsdaten, die nicht frei zugänglich sind, sind Suchmaschinen und soziale Netzwerke – eine Anwendung algorithmischer Entscheidungsprozesse, mit denen die Mehrheit der Internetnutzer täglich konfrontiert ist. Die Strukturierung, Personalisierung und Bewertung von Inhalten erledigen bei sozialen Netzwerken und Suchmaschinen algorithmische Systeme, die als wesentliche Signale die Reaktionen der Nutzer auswerten (Lischka und Stöcker 2017: 15). Diese Reaktionen kann kein anderer Anbieter auswerten, um eigene Empfehlungssysteme zu entwickeln. Die Konzentration der Nutzer bei wenigen Anbietern verbessert ihre Datenlage und verschafft ihnen Vorteile gegenüber neuer Konkurrenz. Im Jahr 2014 vermittelten Google und Facebook

mehr als die Hälfte der Zugriffe auf Webangebote von Medien. Der Anteil steigt seit Jahren kontinuierlich. Insgesamt entfällt inzwischen ein Anteil von fast 75 Prozent des Traffic im Web auf Angebote, die zu Google oder Facebook gehören (Staltz 2017). Das hat für die Forschung gravierende Folgen: 2015 erschien in „Science“ eine Studie zum Nutzungsverhalten von Facebooknutzern (Bakshy, Messing und Adamic 2015). Die drei Autoren waren Facebookangestellte, die Datenbasis ihrer Forschung ist für niemanden außerhalb von Facebook zugänglich.

Eine ähnliche Konzentration diagnostiziert Calo in der gesamten Entwicklung algorithmischer Systeme:

> *„The reality that a handful of large entities (literally, fewer than a human has fingers) possess orders of magnitude more data than anyone else leads to a policy question around data parity. Smaller firms will have trouble entering and competing in the marketplace. Industry research labs will come to far outstrip public labs or universities, to the extend they do not already. Accordingly, cutting-edge AI practitioners will face even greater incentives to enter the private sphere, and ML applications will bend systematically toward the goals of profit-driven companies and not society at large. Companies will possess not only more and better information but a monopoly on its serious analysis“* (Calo 2017: 20).

Konkretisierungen der Idee freier Datensätze müssen aus unserer Sicht drei wesentliche Fragen adressieren:

- Wie können legitime wirtschaftliche Interessen kommerzieller Unternehmen in der Interessenabwägung beachtet werden?
- Wie lassen sich freie Trainingsdatensätze und Datenschutz in Einklang bringen?
- Wie lassen sich unerwünschte Nachfolgenutzungen adressieren?

Ein Vorschlag, um diesen Konzentrationstendenzen entgegenzuwirken: Öffentlich geförderte Forschung sollte produzierte Datensets der Allgemeinheit frei zugänglich machen – „Open Data“ für vielfältige algorithmische Systeme. So ein Vorschlag des National Science and Technology Council in einem Bericht für den US-Präsidenten:

> *„Encouraging the sharing of AI datasets – especially for government-funded research – would likely stimulate innovative AI approaches and solutions. However, technologies are needed to ensure safe sharing of data, since data owners take on risk when sharing their data with the research community. Dataset development and sharing must also follow applicable laws and regulations, and be carried out in an ethical manner“* (National Science and Technology Council 2016: 31).

Forschungsförderung durch die öffentliche Hand und gemeinwohlorientierte Akteure sollten sicherstellen, dass Budget für solche Maßnahmen zu Verfügung steht. Die britische wissenschaftliche Gesellschaft Royal Society empfiehlt:

> *„Research funders should ensure that data handling, including the cost of preparing data and metadata, and associated costs, such as staff, is supported as a key part of research funding, and that researchers are actively encouraged across subject areas to apply for funds to cover this. Research funders should ensure that reviewers and panels assessing grants appreciate the value of such data management“* (Royal Society 2017: 8).

Weiter geht der deutsche Politikberater Philipp Otto: Er plädiert für eine aktive Rolle des Staates über Forschungsförderung hinaus, um auch Daten aus privatwirtschaftlichen Quellen verfügbar zu machen, die zur Erfüllung hoheitlicher Aufgaben nötig sind:

> *„Eine Diskussion der Verwendung von Daten aus privatwirtschaftlichen Bereichen im öffentlichen Interesse würde viele Herausforderungen mit sich bringen, vom Datenschutz bis zum Eigentumsrecht. Dennoch ist die Erörterung dieser Frage auch über die nationalen Grenzen hinaus von Bedeutung. Ein öffentlich kontrollierter bzw. öffentlich nutzbarer europäischer Datenpool kombiniert aus Daten öffentlicher*

> *Institutionen und bestimmten relevanten Daten aus der Privatwirtschaft, die unter klar definierten Konditionen vom Staat genutzt werden dürfen, verspricht zumindest ein attraktives Gedankenexperiment"* (Otto 2017: 30).

Hier ist bei einer Konkretisierung unerlässlich zu fragen, ob dabei überhaupt personenbezogene Datensätze einbeziehbar wären. Ist eine robuste Anonymisierung großer Trainingsdatensätze machbar und wenn ja, wie?

Insbesondere Daten aus dem Kernbereich staatlicher Daseinsvorsorge können genutzt werden, um eine Vielfalt algorithmischer Systeme durch zugängliche Daten zu ermöglichen. Dementsprechend empfiehlt die britische Stiftung Nesta:

> *„Certainly algorithms in fields like welfare to work, health or probation, that are paid for by taxpayers, should be as transparent as possible, and in particular training data should be open since that's what – in many cases – shapes the algorithms"* (Mulgan 2016: 3).

Die Förderung von Trainingsdaten geht über die Sammlung und Bereitstellung von Daten in einer Trainingsdatenallmende hinaus. Damit sie nicht nur Vielfalt, sondern auch die Erklärbarkeit und Überprüfbarkeit algorithmischer Systeme erhöht, müssen weitere Bedingungen erfüllt sein.

4.3.2 Staatliche Nachfrage nach algorithmischen Systemen zur Vielfaltssicherung nutzen

Steckbrief: Öffentliche Stellen entwickeln innovative algorithmische Prozesse in der Daseinsvorsorge

Kerngedanke: In der Daseinsvorsorge ohne kommerzielle Interessen vorbildliche Standards für algorithmische Systeme erproben und setzten

Handlungsfeld: Diversität

Stakeholder: ADM-Entwickler, ADM-Betreiber

Durchsetzender Akteur: Staat

Instrumente: Beschaffung, Entwicklung, Institutionen

Status: Idee

Wie das Beispiel der Biobanken zeigt, gehört zur Vielfalt von Systemen und Betreibern auch eine aktiv gestaltende Rolle der öffentlichen Hand. In den Vereinigten Staaten, aber auch in Deutschland setzen unterschiedliche öffentliche Institutionen algorithmische Systeme zum Vorbereiten oder Treffen von Entscheidungen ein (vgl. Lischka und Klingel 2017; Rohde 2017). Die hier bei der Entwicklung, Implementierung und dem Einsatz demonstrierten Standards genügen nicht dem in diesem Papier formulierten Anspruch an gesellschaftliche Angemessenheit und Erklärbarkeit. Dafür gibt es im Einzelnen unterschiedliche Gründe, doch übergreifend sind bei vielen Fallbeispielen folgende Mängel erkennbar:

- Es fehlt beim staatlichen Einsatz an verbindlichen Standards und Prozessen für ADM-Systeme (z. B. im Hinblick auf Angemessenheit und Erklärbarkeit). Wie sind beispielsweise Prozesse zu dokumentieren?
- Es fehlt beim staatlichen Einsatz an Fachkompetenz für die Gestaltung, Implementierung und Bewertung von algorithmischen Systemen.
- Es fehlt beim staatlichen Einsatz von algorithmischen Entscheidungssystemen der über den jeweiligen Einsatzzweck hinausgehende Anspruch, exemplarische Lösungen zu entwickeln, die als Vorbild für andere Anwendungen dienen könnten.

Eine von der der Obama-Regierung eingesetzte Expertenkommission empfahl der Exekutive, die eigene Rolle bei der Beschaffung, Entwicklung und dem Einsatz von algorithmischen Systemen zum Gestalten einer positiven Ordnung zu nutzen. Dazu gehören:

- **Offene Software und Standards fördern**: „To help support a continued high level of innovation in this area, the U.S. government can boost efforts in the development, support, and use of open AI technologies. Particularly beneficial would be open resources that use standardized or open formats and open standards for representing semantic information, including domain ontologies when available. Government may also encourage greater adoption of open AI resources by accelerating the use of open AI technologies within the government itself, and thus help to maintain a low barrier to entry for innovators. Whenever possible, government should contribute algorithms and software to open source projects" (National Science and Technology Council 2016: 32).
- **Standards und Prozesse für den staatlichen Einsatz entwickeln**: „Agencies should work together to develop and share standards and best practices around the use of AI in government operations. Agencies should ensure that Federal employee training programs include relevant AI opportunities" (Executive Office of the President et al. 2016: 16).
- **Kompetenz zur Entwicklung, Implementierung und Bewertung** algorithmischer Systeme in einer übergeordneten Agentur konzentrieren: „The Federal Government should explore ways to improve the capacity of key agencies to apply AI to their mission" (ebd.).

Die Vorschläge greifen ineinander: Den Aufbau von Kompetenzen zur Entwicklung, Implementierung und Bewertung algorithmischer Systeme setzt ihre Nachvollziehbarkeit voraus und eigene Expertise. Diese wird durch freie Software und Standards ermöglicht. Sie bildet die Grundlage für informierte Standardentwicklung und den Austausch von Best Practices zwischen den Behörden.

Steckbrief: Qualitätsanforderungen an algorithmische Systeme in staatlicher Softwarebeschaffung reglementieren

Kerngedanke: In öffentlichen Ausschreibungen für ADM-Systeme Standards für Überprüfbarkeit, Folgenabschätzung, offene Standards verpflichtend aufnehmen

Handlungsfeld: Diversität

Stakeholder: staatliche Stellen, ADM-Entwickler, ADM-Betreiber

Durchsetzender Akteur: Staat

Instrument: Ausschreibungsrecht

Status: Idee, mögliche Vorbilder in anderen Bereichen

Bei der staatlichen Beschaffung gehören Sozial- und Umweltstandards nicht selten zu den Vergabekriterien. Die Entwicklung und Einhaltung solcher Standards wird zum Teil staatlich gefördert, zum Beispiel auf kommunaler Ebene vom bundesweiten Netzwerk zur fairen Beschaffung, das vom Bundesentwicklungsministerium finanziert wird. Das Instrument bietet sich zur langfristigen Unterstützung der Diversität von algorithmischen Entscheidungssystemen an – sowohl mit Blick auf die Vielfalt der Systeme wie auch auf die Vielfalt der Sektoren.

Calo schlägt vor:

> *„In addition, and sometimes less well recognized, the government can influence policy through what it decides to purchase. States are capable of exerting considerable market pressures. Thus, policymakers at all levels ought to be thinking about the qualities and characteristics of the AI-enabled products government will purchase and the companies that create them. Policymakers can also use contract to help ensure best practice around privacy, security, and other values. This can in turn move the entire market toward more responsible practice and benefit society overall"* (Calo 2017: 24).

Die Idee, staatliche Nachfragemacht zur gemeinwohlorientierten Gestaltung zu nutzen, und die auf anderen Gebieten dabei etablierten Verfahren sollten Inspiration für Beschaffungsstandards für algorithmische Systeme sein. Darüber hinaus können Anforderungen an Erklärbarkeit, Angemessenheit und Diversität von Systemen die Entwicklung entsprechender Praktiken und Werkzeuge fördern. So könnten staatliche Stellen bei Ausschreibungen von ADM-Verfahren Instrumente zur Stakeholderpartizipation (vgl. Kapitel 4.1.4) oder Umsetzungsprüfung (vgl.

Kapitel 4.2.1) verpflichtend machen und Anforderungen an die Nutzung und Förderung offener Standards und Software und die Verfügbarkeit von Trainingsdaten (vgl. Kapitel 4.3.1). stellen.

Darüber hinaus bietet staatliche Nachfrage die Option, die Entwicklung und Anwendung algorithmischer Entscheidungssysteme einem Monitoring zu unterziehen, Mindeststandards in der Entscheidungsforensik zu setzen, die Auszeichnung der Trainingsdaten und deren Dokumentation voranzubringen und sie zu einem Gegenstand von Impact Assessments zu machen. Ein Beispiel für die Umsetzung solcher Empfehlungen ist die Gesetzesnovelle der Stadt New York, die nach gravierenden Problemen von algorithmischen Entscheidungssystemen in der öffentlichen Verwaltung verabschiedet wurde. Sie sieht unter anderem vor, dass eine Expertengruppe Kriterien für die Auswahl von Systemen entwickelt, die zukünftig zum Einsatz kommen (The New York City Council 2018).

4.3.3 Gemeinwohlorientierte Entwicklung algorithmischer Prozesse fördern

Steckbrief: Förderprogramme und Standards für gemeinwohlorientierte Entwicklung aufsetzen

Kerngedanke: Nicht kommerzielle, aber gesellschaftlich motivierte ADM-Entwicklung fördern

Handlungsfeld: Diversität

Stakeholder: Forschung, ADM-Entwickler, ADM-Betreiber, Nichtregierungsorganisationen

Durchsetzende Akteure: staatliche Nichtregierungsorganisationen

Instrumente: Förderprogramme

Status: Idee, mögliche Vorbilder in anderen Bereichen

Forschungs- und Entwicklungsförderung ist ein weiteres Instrument, um durch staatliche Investitionen Technikentwicklung gemeinwohlförderlich zu gestalten. Calo sieht Handlungsbedarf bei der Grundlagenforschung und den Untersuchungen zur gesellschaftlichen Einbettung algorithmischer Systeme. In Kapitel 4.2 genannte Instrumente wie standardisierte Folgenabschätzungen oder professionsethische Kodizes müssen entwickelt und erprobt werden. Sie bieten konkrete Ansatzpunkte für angewandte Forschung zur Einbettung von algorithmischen Systemen in gesellschaftliche Kontexte:

> *„Investment opportunities include not only basic AI research, which advance the state of computer science and help ensure the United States remains globally competitive, but also support of social scientific research into AI's impacts on society. Policymakers can be strategic about where funds are committed and emphasize, for example, projects with an interdisciplinary research agenda and a vision for the public good"* (Calo 2017: 24).

Zu prüfen und konkretisieren sind bei einer Ausarbeitung von Ideen zur gemeinwohlorientierten Forschungsförderung diese Aspekte:

- Standards zu Transparenz und Nachvollziehbarkeit bei Studien und Projekten ausarbeiten, durchsetzen und zum Beispiel zu einem Kriterium für Förderungsfähigkeit machen.
- Open-Data- und Open-Source-Standards definieren und zu einem Kriterium für Förderungsfähigkeit machen.
- Die Wirkung der Forschung auf die Praxis zu einem Kriterium für Förderungsfähigkeit machen.

Die Förderung gemeinwohlorientierter algorithmischer Systeme sollte sich nicht allein auf etablierte Forschungsinstitutionen konzentrieren. Projekte unabhängiger Freiwilliger im Open-Source-Bereich oder in zivilgesellschaftlichen Initiativen stellen ebenfalls wertvolle Quellen von Expertise dar (Kapitel 4.2.6). Ein positives Beispiel dafür ist der Prototype Fund: Das von der Open Knowledge Foundation verantwortete und vom Bundesministerium für Bildung und Forschung finanzierte öffentliche Förderprogramm unterstützt „gemeinnützige Softwareprojekte in den Bereichen Civic Tech, Data Literacy und Datensicherheit". Die geförderten Projekte sollen bis zur ersten Demoentwicklung gebracht werden – mit finanzieller Unterstützung und Coachingangeboten (Open Knowledge Foundation Deutschland o. J.).

Der Prototype Fund ist ein Beispiel für die Förderung von Vielfalt im Start-up-Bereich. Um vielfältige Betreibermodelle algorithmischer Systeme zu fördern, kann staatliche Gründerförderung auch stärker genossenschaftliche und gemeinnützige Organisationsformen (wie gemeinnützige GmbHs) in den Blick nehmen und diese nicht nur finanziell fördern, sondern durch Organisations- und Aufbauberatung unterstützen. In den Vereinigten Staaten zeigen die Wikimedia Foundation und die Mozilla Foundation, dass Betreibermodelle des Dritten Sektors die Angebotsvielfalt algorithmischer Systeme vergrößern und auch langfristig sichern können.

4.4 Übergreifende Rahmenbedingungen für teilhabeförderlichen ADM-Einsatz schaffen

Herausforderungen und Lösungsansätze im Bereich algorithmischer Entscheidungsfindung sind vielfältig. Bisher wurden unterschiedliche Handlungsoptionen für staatliche, wirtschaftliche, wissenschaftliche und zivilgesellschaftliche Akteure aufgezeigt. Ergänzend werden nun übergreifende Aufgaben für den Staat diskutiert, um die einzelnen Initiativen zu fördern und zu koordinieren.

4.4.1 Gesetzlichen Rahmen auf Anpassungsbedarf prüfen

Steckbrief: Effektivität der Rechtsetzung und Durchsetzung bei ADM analysieren
Kerngedanke: Regelungslücken und Durchsetzungslücken existierender Regulierung identifizieren
Handlungsfelder: übergreifende Rahmenbedingungen, Zielsetzung prüfen, Umsetzung prüfen
Stakeholder: Exekutive, Legislative, ADM-Entwickler, ADM-Betreiber, Nichtregierungsorganisationen
Durchsetzende Akteure: Staat (Rechtsetzung, Analyse), Forschung (Analyse), Nichtregierungsorganisationen (Analyse)
Instrumente: rechtliche Analyse
Status: erste Ansätze

Zwei Kernfragen sind mit Blick auf den rechtlichen Rahmen für algorithmische Systeme zu beantworten:

- Bestehen **Regelungslücken im Recht**?
- Bestehen **Lücken bei der Durchsetzung geltenden Rechts**?

Eine umfassende, systematische Analyse dieser Fragen steht bislang aus. Wie bei der Frage nach effektiver öffentlicher Aufsicht (vgl. Kapitel 4.2.5) ist eine Herausforderung, zwei Aspekte in Einklang zu bringen:

- die **Besonderheiten algorithmischer Systeme** (vgl. Kapitel 3) sowie
- die **Besonderheiten der Sektoren**, in denen algorithmische Systeme zum Einsatz kommen.

Für die vorrangige Orientierung am Einsatzumfeld plädiert beispielsweise Jaume-Palasí:

> *„Algorithmen gibt es in allen Bereichen (Finanzen, Gesundheit, Wirtschaft, Bildung, Verkehr, Industrie, Kommunikation, etc. etc.). In einigen dieser Bereiche gibt es bereits Kontrollinstanzen und -mechanismen, die den für eine Beurteilung notwendigen Kontext viel besser einschätzen können und gegebenenfalls eher einer Anpassung oder Neujustierung bereits bestehender Gesetze bedürfen (beispielsweise etwa die deutsche Finanzmarktaufsicht (BaFin), das Bundesinstitut für Arzneimittel und Medizinprodukte oder der TÜV für Fahrzeuge)“* (Jaume-Palasí 2017: 1).

Sehr wahrscheinlich wird eine Analyse des rechtlichen Rahmens zum Ergebnis kommen, dass sowohl Regelungslücken als auch Durchsetzunglücken bestehen – an unterschiedlichen Stellen. Erste Hinweise auf **Regelungslücken** im Recht liefert beispielsweise Martinis Analyse der Wirksamkeit des Allgemeinen Gleichbehandlungsgesetzes (AGG) bei algorithmischen Systemen:

> *„Die regulatorische Zielsetzung, Diskriminierungsrisiken algorithmenbasierter Verfahren zu begrenzen, geht mit der Schutzmission des Allgemeinen Gleichbehandlungsgesetzes (AGG) Hand in Hand: Beide sollen die Benachteiligung diskriminierungsgefährdeter Menschen – typischerweise Minderheiten – verhindern. Das AGG schließt zwar softwarebasierte Verfahren schon heute nicht von seinem Anwendungsbereich aus; es ist technologieneutral konzipiert. Zugleich ist es aber auf einen begrenzten Kanon von Lebensbereichen limitiert, namentlich auf Arbeitsverhältnisse, Bildung und Sozialleistungen sowie Leistungen, die der allgemeinen Öffentlichkeit zur Verfügung stehen (§§ 2 und 19 AGG). Für Verträge zwischen Privaten außerhalb des Arbeitsrechts gilt das Gesetz nur bei sog. Massengeschäften und Versicherungen – von der Diskotür bis zur Krankenversicherung, ob analog oder digital. Zahlreiche spezialisierte Anwendungsfelder softwarebasierter Verfahren erfasst das AGG demgegenüber nicht. De lege ferenda ist eine Ergänzung des Katalogs der Anwendungsfälle des § 2 Abs. 1 AGG um eine Nr. 9 für Ungleichbehandlungen zwischen Privaten erwägenswert, die auf einer algorithmenbasierten Datenauswertung oder einem automatisierten Entscheidungsverfahren beruhen“* (Martini 2017: 1021).

Verbandsklagerechte der Verbraucherschutzverbände, der Antidiskriminierungsverbände und der Antidiskriminierungsstelle des Bundes existieren als Ideen, nicht als gesetzliche Grundlage (vgl. Kapitel 4.2.4.) – eine Regelungslücke.

Eine besondere Form der Regelungslücke sind zu verbietende algorithmische Verfahren. **Gesetzliche Verbote** (ggf. mit Erlaubnisvorbehalten) lassen sich mit strukturellen Besonderheiten algorithmischer Systeme und/oder sektoralen Besonderheiten begründen. Wenn zum Beispiel bei automatisiert lernenden Systemen ein Mindestmaß an Überprüfbarkeit und Nachvollziehbarkeit (vgl. Kapitel 4.2) von einzelnen Outputs nicht gewährleistet ist, fordern einige Experten Anwendungsverbote in besonders teilhaberelevanten Bereichen, etwa in der Rechtsprechung, Gesundheitsversorgung, Bildung und sozialen Sicherung (Campolo et al. 2017; Eckersley, Gillula und Williams o. J.).

Andere Kriterien für Verbote könnten Sicherheitsrisiken aufgrund des Einsatzfeldes und der Konstruktionsweise der Systeme (z. B. mangelnde Robustheit) sein. Oder Einsatzfelder und Optimierungsziele, die übergeordneten gesellschaftlichen Prinzipien widersprechen. Wo die Gesellschaft sich zum Beispiel für Solidarität und Vergemeinschaftung von Risiken entschieden hat – etwa in der Sozialversicherung – dürfen algorithmische Prozesse diese gezielt kollektivierten Risiken nicht reindividualisieren (Lischka und Klingel 2017: 7).

Als erste Hinweise auf **Lücken bei der Durchsetzung** geltenden Rechts haben wir in dieser Analyse bereits folgende Punkte thematisiert:

- Methoden zur Umsetzungsprüfung (vgl. Kapitel 4.2.1)
- Gesetzlicher Rahmen für Überprüfbarkeit (vgl. Kapitel 4.2.3)

Darüber hinaus sind grundlegende Fragen der **Rechtsdurchsetzung** zu beantworten. Zum Beispiel:

- Wie lassen sich teilweise fortwährend in hoher Frequenz veränderte Optimierungsziele eines algorithmischen Systems nachvollziehen?
- Wie können diese Veränderungen rückblickend analysiert werden (z. B. durch detaillierte Entscheidungsforensik in Auditlogs)?
- Wie lässt sich veränderter Output eines algorithmischen Systems rückblickend nachvollziehen?
- Wie lassen sich rückblickend kausale Verbindungen zwischen Veränderungen bei Optimierungszielen, Dateninput und Systemoutput ziehen?
- Braucht es zur Beantwortung dieser Fragen ggf. Standards und Anforderungen an gerichtsfeste Dokumentation der Optimierungsziele, der Umsetzung, der Outputs und der Inputs?

Diese Herausforderungen der Rechtsdurchsetzung konkretisiert Martini beispielhaft an algorithmischen Systemen, die automatisiert lernen:

> *„Für komplexe Softwareanwendungen in persönlichkeitssensiblen Anwendungsfeldern resultiert die Notwendigkeit kontinuierlicher Kontrolle schon daraus, dass sie ihr Verhalten im Laufe ihres Einsatzes häufig wie ein Chamäleon verändern – sei es durch Updates, sei es aufgrund eines maschinellen Lernverfahrens. Ein obsiegendes Urteil gegen eine diskriminierende softwarebasierte Entscheidung, das zwei Jahre nach der Rechtsverletzung Rechtskraft erlangt, ist dann in seiner Aussagekraft längst überholt“* (Martini 2017: 1021).

Haftungsfragen zeigen beispielhaft, welche Herausforderungen bei der **Rechtsdurchsetzung** zu lösen sind. Bei selbstfahrenden Autos bildete sich in Deutschland mit der im Juli 2017 geänderten Straßenverkehrsordnung ein Haftungsregime aus, dessen Bewährung aussteht: Es sieht vor, dass auch bei Einsatz von Computern die letzte Verantwortung beim Menschen liegt. Der Fahrer muss trotz automatisierter Fahrfunktion wahrnehmungsbereit bleiben und die Fahrzeugführung übernehmen, wenn dies durch den Autopiloten kommuniziert wird oder die adäquate Fahrfunktion durch den Autopiloten dem Fahrer nicht mehr gewährleistet scheint. Es bestätigt damit aktuell das Prinzip der Halterhaftung, mit entsprechenden Versicherungspflichten. Bei einem Unfall soll eine Art Black Box (sic!) dabei helfen aufzuklären, ob es sich um technisches oder menschliches Versagen handelt. Sie zeichnet die wesentlichen Daten der Fahrt auf (Bundesregierung 2017; forum 2017). Bei der Durchsetzung sind Lösungen zur **Dokumentation der Abläufe** und des **Zusammenwirkens unterschiedlicher Komponenten** in algorithmischen Systemen gefragt. Erste Vorschläge:

- Systembetreiber auf **Mindeststandards der Entscheidungsforensik** verpflichten (Citron 2008: 1301 ff.; Tutt 2016).
- **Versicherungspflichten** an die tatsächliche Gestaltung algorithmischer Systeme anknüpfen (ACM 2017; Shneiderman 2016).

4.4.2 Staatliche Regulierungskompetenz stärken

Steckbrief: Zentrale Agentur für algorithmische Systeme und sektorspezifische Regulierungslösungen entwickeln
Kerngedanke: Agentur für algorithmische Systeme gründen
Handlungsfelder: übergreifende Rahmenbedingungen, Zielsetzung prüfen, Umsetzung prüfen
Stakeholder: Staat
Durchsetzender Akteur: Staat
Instrument: Institution
Status: Idee

Die Rolle des Staates als Entwickler und Auftraggeber einzelner algorithmischer Systeme in der Daseinsvorsorge wurde im Kontext der Vielfaltssicherung thematisiert (vgl. Kapitel 4.3.2). Doch staatliche Kompetenzen müssen bei algorithmischen Prozessen den gesamten Rahmen umfassen – nicht nur Beauftragung, sondern Bewertung, Kontrolle und Regulierung.

Um der Komplexität algorithmischer Entscheidungen in der Gesellschaft gerecht zu werden, schlagen unterschiedliche Autoren vor, staatliche Kompetenzen zentralisiert aufzubauen oder entsprechende Kompetenzen bei existierenden Institutionen anzusiedeln. Bei allen sektorspezifischen Besonderheiten gibt es übergeordnete Aufgaben, die bei einer zentralen **Agentur für algorithmische Systeme** anzusiedeln wären, insbesondere die Beobachtung und Analyse dieser Punkte:

- technologische Entwicklung und konkrete Anwendungsfelder
- übergeordnete, systemische Risiken
- Klassifikation von algorithmischen Entscheidungssystemen, beispielsweise in Abhängigkeit ihrer Funktionsweise und Komplexität, dem Einsatzgebiet und ihrer Risiken
- Identifizierung von algorithmischen Entscheidungssystemen, die einer Regulierung im Sinne einer hoheitlichen Aufsicht oder Kontrolle in Form von Auditing, Zertifizierung oder einem Verbot unterliegen sollten

- Aufsicht und Kontrolle über die sachgerechte Verwendung von Daten
- Entwicklung von Sicherheits-, Forschungs- und Anwendungsstandards
- Regelung von Haftbarkeiten und Prüfungsverfahren (Tutt 2016)

Die meisten Autoren verorten eine solche zentrale Institution im Regelfall als Beratungsinstanz für Legislative, Exekutive und Judikative (Calo 2017; Cave 2017; Executive Office of the President et al. 2016; Mulgan 2016). Tutts Vorschläge gehen darüber hinaus: Er befürwortet die Einrichtung einer zentralen Behörde, die sowohl die weiche Regulierung wie etwa Transparenzvorschriften und Standardsetzung zur Aufgabe hat als auch die Zertifizierung und Zulassung von Systemen übernimmt:

> „*The rise of increasingly complex algorithms calls for critical thought about how best to prevent, deter, and compensate for the harms that they cause. This paper argues that the criminal law and tort regulatory systems will prove no match for the difficult regulatory puzzles algorithms pose. Algorithmic regulation will require federal uniformity, expert judgment, political independence, and pre-market review to prevent – without stifling innovation – the introduction of unacceptably dangerous algorithms into the market* (Tutt 2016: 1).

Eine **Agentur für algorithmische Systeme** mit dieser Ausrichtung erinnert an das 2016 vom Bundeswirtschaftsministerium in die Diskussion eingebrachte Konzept einer Bundesdigitalagentur zwecks „Bündelung von Kompetenzen, Unterstützung der politischen digitalen Agenda, nachhaltigem Aufbau von Digitalisierungskompetenz“ (Bundesministerium für Wirtschaft und Energie 2016: 56). Analog dazu lassen sich die Ziele einer Agentur für algorithmische Systeme beschreiben als: Aufbau und Anwendung rechtlicher, technischer, gesellschaftlicher Kompetenz zu gemeinwohlorientierter Gestaltung algorithmischer Systeme.

Auf der anderen Seite bedarf es insbesondere im Bereich hoheitlicher und öffentlicher Aufgaben auch des Ausbaus dieser **Kompetenzen bei existierenden Institutionen**, so Stone et al. (2016). Dies erfordert aber Kompetenz im Umgang mit algorithmischen Prozessen als Querschnittsthema. Ziel ist hier einerseits das Verständnis für die Wechselwirkungen zwischen algorithmischer Entscheidungsfindung, Politikprogrammen und gesellschaftlichen Zielen. Da algorithmische Prozesse nur im Einsatzkontext beurteilt werden können, ist Domänenfachwissen unabdingbar. Bewertungen sind kontextabhängig, es braucht zum Beispiel Wissen um Medikamentenwirkung bei der Prüfung von algorithmischen Systemen in der Pharmakologie.

Kompetenzausbau an den fachlich erfahrenen Stellen bietet die Möglichkeit, dass algorithmische Entscheidungssysteme eng mit politischen Prämissen verzahnt werden und spezifische Verwaltungsexpertise mit in das Design algorithmischer Entscheidungsprozesse einfließen kann (Executive Office of the President et al. 2016; Stone et al. 2016).

Staatliche Regulierung hinkt der technischen Entwicklung im Regelfall hinterher. Was auch daran liegen kann, dass die Fragestellungen andere sind. Staatliche Expertise ist nicht auf die Auswirkungen in einem Geschäftsfeld mit relativ klar umrissenen Optimierungszielen beschränkt, sondern hat das Gemeinwesen und alle Partikularinteressen im Blick. Demgegenüber stellen Unternehmen Experten auf neuen Feldern oft eher ein und bieten zudem attraktive Aufstiegs- und Verdienstmöglichkeiten (Schuetze 2018). Daher stellt der Ausbau staatlicher Kompetenzen sowohl auf übergeordneter Ebene sowie in den Einzelbereichen eine große Herausforderung dar. Gleichzeitig bietet er natürlich die Chance, Souveränität auf dem Gebiet algorithmischer Entscheidungssysteme zu bewahren und auszubauen sowie den digitalen Wandel zu gestalten (Calo 2017: 23).

4.4.3 Individuelle Sensibilisierung und Kompetenz steigern

Steckbrief

Kerngedanke: Die Gestaltungskompetenz potenziell Betroffener im Umgang mit algorithmischen Systemen stärken

Handlungsfelder: übergreifende Rahmenbedingungen, Zielsetzung prüfen, Umsetzung prüfen

Stakeholder: Bürger

Durchsetzende Akteure: Staat, Wirtschaft, Nichtregierungsorganisationen

Instrumente: Bildungskonzepte

Status: erste Ideen

Stakeholderpartizipation, Widerspruchsverfahren, Auskunftsrechte, zivilgesellschaftliche Wächter – solche Lösungsvorschläge eint, dass sie für die Umsetzung auf das Mitwirken der Bürger angewiesen ist. Von algorithmischen Entscheidungen betroffene Menschen müssen etwa Auskunft verlangen, Informationen an Wächterorganisationen weiterleiten oder Widerspruchsrechte in Anspruch nehmen können. Das setzt ein Grundwissen darüber voraus, wo algorithmische Entscheidungen im Einsatz sind, welche Chancen und Risiken der Einsatz hat und wie man als (potenziell) Betroffener Einfluss auf dessen Gestaltung und Einsatz nehmen kann. Dabei darf die Verantwortung nicht beim Individuum allein verortet werden. Unabhängig vom Bildungshintergrund und dem Vorwissen um algorithmische Prozesse muss jeder Bürger in der Lage sein, sich gegen fragwürdige Prozesse zu wehren. Auch, indem er qualifizierte Hilfe von Institutionen zurate zieht. Individuelle Gestaltungskompetenz muss mit der Stärkung und ggf. dem Aufbau von Institutionen einhergehen, die Individuelle Gestaltungskompetenz fördern, unterstützen, ergänzen und zum Teil – wo nötig – ausgleichen.

Neben staatlicher Gestaltungskompetenz und professionsethischen Standards braucht es also auch Kompetenz in der Bevölkerung im Umgang mit algorithmischen Systemen. Die Royal Society definiert diese Kompetenz als:

> *„(…) a basic grounding in what machine learning is, and what it does, will be necessary in order to grasp, at a basic level, how our data is being used, and what this means for the information presented to us by machine learning systems“* (Royal Society 2017: 63).

Ein erster Schritt wäre es, den Kenntnisstand zu erfassen. Bislang wurde nicht ermittelt, wie es um das Wissen über den Einsatz von algorithmischen Systemen und das Verständnis ihrer Funktionsweise in der Bevölkerung in Deutschland bestellt ist. Zunächst wäre zu konkretisieren, wie sich in diesem Bereich Kompetenz äußert und wie sie erfasst werden kann. In den Vereinigten Staaten wird diese Idee unter anderem unter der Bezeichnung „algorithmic literacy“ diskutiert, allerdings fehlt bislang die Operationalisierung:

> *„This group also discussed algorithmic literacy, in terms of reading, writing and making algorithms. This solution is a response to a perception of growing ‚illiteracy‘ and inequality in access to and control of algorithmic mechanisms. Algorithmic literacy programs are, in general, designed to enable more individuals to impact information flows and perceive when or if they or others are being marginalized“* (Caplan, Reed und Mateescu 2016: 8).

Bei der Frage, wie Kompetenz auf diesen Gebiet aufgebaut werden kann, liegt der Ruf nach Schule und Universität nahe – zu dem sich die Royal Society (2017: 63) folgendermaßen äußert:

> *„If introduced at primary or secondary school, a basic understanding of key concepts in machine learning can help with navigating this world, and encourage further uptake of data science subjects“* (Royal Society 2017: 63).

Offen ist allerdings, wie die Kompetenzen der Mehrheit der Bevölkerung gestärkt werden können, die nicht mehr eine Schule besuchen. Zu möglichen Akteuren zählen Volkshochschulen, Verbraucherschutzorganisationen, Datenschutzbehörden, Stiftungen. Es können aber auch zivilgesellschaftliche Wächterorganisationen sein, die Informationen für die Öffentlichkeit aufbereiten.

Ein Vorschlag zur Darreichungs- statt zur Organisationsform zum Kompetenzaufbau kommt von den Juristen Danielle Citron und Frank Pasquale, die für eine Veranschaulichung der Entscheidungslogik algorithmischer Systeme in interaktiven Modellen plädieren. Sie zeigen am Beispiel von Kreditscoring-Systemen für Simulationen, an denen Menschen ausprobieren können, welchen Output die Systeme bei verschiedenen Inputwerten liefern:

> *„Another approach would be to give consumers the chance to see what happens to their score with different hypothetical alterations of their credit histories. (…) To make it more concrete, picture a consumer who is facing a dilemma. She sees on her credit report that she has a bill that is thirty days overdue. She could secure a payday loan to pay the bill, but she'd face a usurious interest rate if she takes that option. She can probably earn enough money working overtime to pay the bill herself in forty days. Software could give her an idea of the relative merits of either course. If her score dropped by 100 points when a bill went unpaid for a total of sixty days, she would be much more likely to opt for the payday loan than if a mere five points were deducted for that term of delinquency“* (Citron und Pasquale 2014: 29 f.).

Auch für diesen Vorschlag fehlen bislang konkrete Konzepte zur Umsetzung. Die ersten zu konkretisierenden Punkte sind:

- Woher kommt die Grundlage für die Modellierung der Entscheidungssysteme?
- Wie können Funktionsweisen so veranschaulicht werden, dass Menschen ohne Vorwissen sich auf die interaktiven Modelle einlassen?

Einige der wenigen Beispiele für solche interaktiven Ansätze zur Vermittlung: Der FICO-Kreditrating-Simulator (Free Credit Scores Estimator from myFICO o. J.). Oder die Anwendung Justice.exe, die Informatikstudenten der University of Utah im Rahmen eines Seminars zu algorithmischer Entscheidungsfindung entwickelt haben (University of Utah Honors 2017). Die Software versetzt den Anwender in die Rolle eines Richters. Ihm werden kurz die Taten der Angeklagten geschildert sowie ihre Vorstrafen und ihr sozio-demographischer Hintergrund. Der Anwender entscheidet, ob er die Höchststrafe oder die Mindeststrafe vergibt. Diese Entscheidungen trainieren ein Modell, das nach 50 beurteilten Fällen zeigt, welche Faktoren es als starke Signale für Höchststrafen – Vorstrafen? Hautfarbe? Geschlecht? Alter? Familienstand? Bildung? – erlernt hat und entsprechend erste Prognosen geben kann. Binnen fünf Minuten zeigt dieses auf fiktiven Daten basierende Programm, wie maschinelles Lernen funktioniert und welche Rolle dabei Muster menschlicher Entscheidungen spielen. Unabhängig davon, in welchem organisatorischen Rahmen die notwendige Kompetenzvermittlung stattfindet, sollten die Entwicklung, Umsetzung und der Einsatz solcher Software vorangetrieben werden.

Abbildung 4:

(Quelle: eigene Darstellung)

5 Zusammenfassung und Fazit: Was nun zu tun ist

Menschliches Handeln und menschliche Entscheidungen sind fehlbar. Menschen diskriminieren oft unbewusst, können Komplexität in vielen Situationen nur schwer bewältigen und entscheiden inkonsistent. **Algorithmische Entscheidungsprozesse** können helfen, einige dieser Fehler zu erkennen, auszugleichen und Entscheidungsprozesse **konsistenter** und dadurch womöglich fairer zu machen. Diese Korrekturfunktion ist ein wesentliches Argument für den Einsatz algorithmischer Systeme. So kann zum Beispiel die Ungleichbehandlung von Bewerbern anhand von gesellschaftlich unangemessenen Merkmalen (z. B. ausländisch klingender Name) in algorithmischen Systemen konsistent ausgeschlossen werden. Menschen tendieren nachweislich dazu, solche ungeeigneten Merkmale bei der Bewerberauswahl zu berücksichtigen.

Zudem können algorithmische Systeme die Analyse und Entscheidungsfindung um **Effektivität** in einer neuen Qualität bereichern. Sie können Datenreichtum in einigen Fällen **effizienter** (schneller, ggf. kostengünstiger) und in anderen **effektiver** (wirksamer, nutzbringender) bewältigen als Menschen. Zum Beispiel: Die Crawler[16] von Suchmaschinen analysieren fortwährend die Beziehungen von Milliarden Websites. Software bereitet radiologische Aufnahmen für menschliche Analysen auf, zählt beispielsweise Wirbel und ermöglicht es, die Explosion an Aufnahmen zu bewältigen (Harvey 2018). Software überwacht auf Intensivstationen fortwährend Veränderungen mehrerer Vitalparameter und ihr Zusammenwirken bei allen Patienten (Briseno 2018). In keinem dieser Beispiele könnten Menschen ohne Unterstützung durch algorithmische Systeme unter denselben Rahmenbedingungen Vergleichbares leisten.

Mehr auswertbare Daten und neue Analyseverfahren bieten die Chance, **neue Erkenntnisse** über den Einzelnen und die Gesellschaft zu gewinnen. Diese können zu **qualitativ besseren Entscheidungen** beitragen, etwa durch die Berücksichtigung von mehr oder besseren Daten für einen Einzelfall. Das ermöglicht eine stärkere Personalisierung von Verfahren und einen besseren Umgang mit Komplexität, etwa bei der Verteilung von Ressourcen.

Damit diese Chancen verwirklicht werden und dem Gemeinwohl dienen, müssen Staat, Wirtschaft und Zivilgesellschaft die Entwicklung gestalten. Denn wir haben gesehen: Auch algorithmische Entscheidungssysteme sind fehlbar (vgl. Kapitel 2.2). Handelsbedarf besteht insbesondere in vier Bereichen:

- gesellschaftliche Angemessenheit der Optimierungsziele von algorithmischen Systemen
- Umsetzung der Ziele bei Aufbau und Einsatz der Systeme (z. B. bei der Operationalisierung, Datengrundlage oder Einbettung in den gesellschaftlichen Kontext)
- Vielfalt an Systemen und Betreibermodellen in bestimmten Einsatzbereichen
- übergreifende Rahmenbedingungen wie Gesetzgebung, staatliche und individuelle Gestaltungskompetenz

Dieses Arbeitspapier bietet im Kernkapitel 4 einen Überblick der Lösungsansätze, gegliedert nach diesen vier Handlungsfeldern (vgl. Abbildung 4). Einige dieser Gestaltungsansätze sind im Folgenden kurz zusammengefasst. Dabei kommt diesen Aspekten eine besondere Bedeutung zu:

- **Rechtslage**: Inwieweit beziehen sich Lösungsvorschläge auf die aktuelle Gesetzgebung? Sind sie durch diese gedeckt? Besteht weiterer Handlungsbedarf? Ergeben sich bei der Implementierung von Normen besondere Herausforderungen?
- **Umsetzung**: Wie konkret sind die Lösungsvorschläge? Lassen sie sich einfach implementieren? Gibt es besondere Vorschläge hinsichtlich der institutionellen Verortung und/oder des Designs?

[16] Computerprogramme zum Aufbau und Indexieren von Webseiten, die dafür automatisch das World Wide Web durchsuchen, analysieren und ggf. nach bestimmten Kriterien sortieren.

5.1 Die Ziele und Mittel: Gesellschaftliche Angemessenheit prüfen

Welche Optimierungsziele gesellschaftlich angemessen sind, lässt sich nicht allgemein für alle Anwendungsbereiche festlegen. Die Gesellschaft, Werte und Normen verändern sich fortwährend. Algorithmische Systeme müssen diese Dynamik einbeziehen, statt nur die Vergangenheit zu reproduzieren. Potenziell Betroffene müssen über sie betreffende Systeme informiert sein, um sie mitgestalten zu können. Hierbei kommt der Information aller Beteiligten, die durch verschiedene Akteure und Verfahren realisiert werden kann, eine übergeordnete Rolle zu.

Entwickler, Systembetreiber, Stakeholder: Interessen und Optimierungsziele dokumentieren

Wer algorithmische Entscheidungssysteme entwickelt bzw. die Entwicklung beauftragt, muss unterschiedliche Optimierungsziele und die getroffene Abwägung dokumentieren. Die mit den Zielen verbundenen Interessen und Stakeholdergruppen sollten ebenfalls systematisch erkannt, erfasst, dokumentiert und einbezogen werden.

So machen Entwickler, Systembetreiber und Anwender die getroffenen Wertentscheidungen sichtbar. Als Instrument eignet sich die Entwicklung und Dokumentation einer Interessenmatrix, die die Varianz von Interessen, Stakeholdern und möglichen Optimierungszielen abdeckt.

Entwickler, Systembetreiber, Gesetzgeber: Betroffene über den Einsatz und die Zielstellung von ADM informieren

Um algorithmische Entscheidungen bewerten und ggf. Widerspruchsrechte wahrzunehmen zu können, müssen Betroffene den Einsatz, die Zielsetzung, die angewandten Methoden und die angestrebte Wirkung kennen. In Teilen gibt die EU-Datenschutz-Grundverordnung (DSGVO-EU) im Bereich der vollautomatisierten Entscheidung Normen vor. Teilweise bedarf der Bereich weiterer Regulierung. Zum Umsetzen einer effektiven Transparenz müssen Verfahren und Instrumente entwickeln werden. Denn Informations- und Transparenzpflichten sind die eine Sache, deren effektive Gewährleistung hingegen eine andere. Der **Datenbrief** vom Chaos Computer Club ist ein derartiges Konzept. Die Idee dahinter: Betroffene erhalten vom Systembetreiber regelmäßig Informationen über die zu ihrer Person gespeicherten Daten und die Verknüpfung dieser Daten mit anderen Informationen sowie abgeleitete Profile, Annahmen über Präferenzen oder Bewertungen (z. B. Kundenklassen).

Ein weiteres Konzept sind die **Anwendungserläuterungen** (Counterfactual Explanations). Der Ansatz soll algorithmische Entscheidungen selbstlernender Algorithmen und komplexer Systeme effektiv erklären helfen. Im einfachsten Fall, etwa bei der Kreditbewilligung, könnte so eine Anwendungserläuterung Auskunft dazu geben, wie hoch das Jahreseinkommen eines Antragstellers sein müsste, damit ein abgelehnter Kreditantrag bewilligt würde. In den Worten der Erfinder:

> *„These counterfactual explanations describe the smallest change to the world that would obtain a desirable outcome, or to arrive at a ‚close possible world'"* (Wachter, Mittelstadt und Russell 2017: 1).

Gleichzeitig ist sicherzustellen, dass die Inputparameter korrekt sind – indem die zugrunde liegende Datenbank eine hohe Qualität ohne Verzerrungen aufweist –, verbunden mit einer adäquaten Entscheidungslogik. Die Kontrolle algorithmischer Entscheidungssysteme setzt eine detaillierte Methodenkenntnis voraus, die hohe Anforderungen auch an den Verbraucherschutz stellt, der eine wichtige Wächterfunktion wahrnehmen kann.

Gesetzgeber, Systembetreiber: Erwartete Folgen und Auswirkung reflektieren und dokumentieren

Die bisherigen rechtlichen Normen geben nur bedingt Antworten auf überindividuelle Fragen, zum Beispiel nach der gesellschaftlichen Angemessenheit oder der Vielfalt von algorithmischen Entscheidungssystemen. Die Transparenz von Zielen, Methoden und Ergebnissen algorithmischer Entscheidungssysteme gegenüber der Öffentlichkeit ist ebenso bedeutsam wie die gegenüber Betroffenen, insbesondere mit Blick auf deren Akzeptanz

und Mitgestaltungsmöglichkeiten der Entscheidungssysteme Voraussetzung für Beurteilung, Kritik oder Einflussnahme seitens Betroffener oder ihrer Interessenvertreter ist ein Mindestmaß an Kenntnis über den Einsatz algorithmischer Verfahren.

Hier bieten sogenannte **Beipackzettel** erste Anregungen, wie Systembetreiber verpflichtet werden könnten, ihren Beitrag für Transparenz zu leisten: **Algorithmenfolgeabschätzungen** bzw. **-verträglichkeitsprüfungen** bieten allgemeine Informationen zu den Zielen, Methoden und Ergebnissen von algorithmischer Entscheidungsfindung. Um Diskriminierungseffekten vorzubeugen, könnten diese um Informationen über konkurrierende Systeme im gleichen Anwendungsbereich ergänzt werden. Solche Impact Assessments bieten Systembetreibern die Möglichkeit, zwischen mehreren Systemen auszuwählen und neue Fehlerquellen zu identifizieren (z. B. Datenverzerrung an Schnittstellen zu Datenbanken). Gesetzgeber und gleichermaßen Systembetreiber sind gefragt, Regulierungsoptionen zu entwickeln. Sie können von der Selbstverpflichtung zur Durchführung von Impact Assessments bis hin zur gesetzlichen Verpflichtung unter Sanktionsandrohungen reichen.

Entwickler, (institutionelle) Anwender: Methoden der Stakeholderpartizipation institutionalisieren

Nur, wenn alle relevanten Stakeholder in die Entwicklung von algorithmischen Systemen einbezogen sind, kann die gesellschaftliche Angemessenheit der gesetzten Optimierungsziele reflektiert werden. Die Entwicklung, Implementierung und Anwendung algorithmischer Systeme beinhaltet von Beginn an Wertentscheidungen, etwa für die Datenauswahl, Methodik und Optimierungsziele. Treuhänder, Gutachtergremien und Interessenvertretungen von Betroffenen sind Beispiele gelungener Partizipation von Stakeholdern aus anderen Anwendungsbereichen. Hier gilt es, für den jeweiligen institutionellen Kontext passende Modelle zu entwickeln und zu erproben.

Entwickler, Forschungs- und Ausbildungseinrichtungen: Professionsethik etablieren

Bei der Gestaltung algorithmischer Entscheidungssysteme tragen alle beteiligten Akteure (z. B., aber nicht ausschließlich Mathematiker, Ingenieure, Informatiker u. a.) ein besonderes Maß an Verantwortung. Sie setzen die definierten Ziele in Codes und Prozesse um bzw. leisten ihren Beitrag zur Entwicklung komplexer algorithmischer Entscheidungssysteme. Insbesondere für sie gilt es, die eigene Rolle zu reflektieren und das Bewusstsein für die gesellschaftliche Verantwortung, die sie tragen, zu schärfen. Unterschiedliche Überlegungen zur Etablierung und Durchsetzung einer **Professionsethik** fokussieren auf die Weiterentwicklung und Konkretisierung verbindlicher Standards, Ausbildungsziele und Beratungsgremien. Ziel ist es, prozessbezogene Qualitätsstandards für die Gestaltung algorithmischer Systeme zu kodifizieren, um Mindeststandards – zum Beispiel an Sorgfalt, Erklärbarkeit oder Folgenabschätzung – zu sichern.

5.2 Die Wirkung: Umsetzung von Zielen in algorithmischen Systemen prüfen

Die Überprüfung algorithmischer Entscheidungssysteme, ihrer Funktionalität und Wirkung ist ein zentrales Thema in der öffentlichen Debatte und die Grundlage für die Evaluation und Weiterentwicklung dieser Systeme nach Maßgabe der gesteckten Ziele.

Entwickler, Forscher, Regulierer: Methoden zum Auditing algorithmischer Systeme entwickeln, erproben, institutionalisieren

Zur Überprüfung vieler der heute eingesetzten statischen Systeme werden klassische Formen von **Algorithmenanalysen (Auditings)** herangezogen. Das Augenmerk der Prüfung richtet sich dabei auf die dem Design, der Problemmodellierung und der Implementierung des Systems zugrunde liegende Logik. Diese Verfahren stoßen allerdings bei komplexen, dynamischen arithmetischen Systemen an ihre Grenzen. Insbesondere die Integration von selbstlernenden Algorithmen erfordert neue Formen des Auditings mit dem Ziel, den Input und Output nach Maßgabe der Kenntnis der Datengrundlage zu überprüfen. Dieses aktuell als hoch dynamisch zu bewertende Feld beinhaltet auch die Entwicklung von komplexen **systemerklärenden Technologien**. Wie bei

den Anwendungserläuterungen gilt auch für diese, dass sie sich weder zur Prüfung der Inputparameter noch zur Evaluation gesellschaftlicher Angemessenheit/korrekter Funktionsweise eignen. Der Schwerpunkt liegt hier auf dem Verständnis darüber, wie konkrete Entscheidungen zustande gekommen sind.

Eng verbunden mit der Entwicklung von Auditings ist die Entwicklung von Standards des algorithmischen Entscheidungssystems. Dies betrifft prozedurale Erfordernisse, Dokumentation und Verschlüsselungstechnologien.

Gesetzgeber, Systembetreiber, zivilgesellschaftliche Wächter: Auditingbedarf priorisieren

Weil das Algorithmenauditing ressourcenintensiv ist, ist eine **Klassifikation algorithmischer Entscheidungssysteme** erforderlich, die dabei hilft, den gesellschaftlich relevanten Auditingbedarf zu ermitteln und nach dessen Maßgabe die Ressourcen zu steuern. Diese Priorisierung sollte neben der Art und Komplexität der Systeme auch ihren Anwendungsbereich sowie mögliche Risiken für Betroffene berücksichtigen.

Gesetzgeber, (institutionelle) Anwender, Systembetreiber: Auditing adäquat institutionalisieren

Unterschiedliche institutionelle Lösungen sind denkbar, eine **qualifizierte Kontrolle** bzw. **Transparenz** herzustellen: Zum Schutz berechtigter Interessen seitens der Betreiber und/oder Betroffenen (Datenschutz) können Gutachterkreise, Behörden oder unabhängige Institutionen etabliert werden, die zugleich Vertraulichkeit und die Prüfung algorithmischer Entscheidungssysteme ermöglichen.

Gesetzgeber, Systembetreiber: Qualität und Herkunft der Datenbasis auszeichnen, Korrektur- und Anwendungsstandards schaffen

Neben der Prüfung ganzer algorithmischer Entscheidungssysteme kommt der Prüfung ihrer elementaren Grundlage – Daten – eine entscheidende Bedeutung zu. Es gilt, die **Korrektheit, Aktualität und Repräsentativität von Datensets** zu gewährleisten.

Denn nur in ihrer Gesamtheit stellen die Daten eine adäquate Basis dar, auf der einerseits die Versprechen lernender Systeme verwirklicht und andererseits Risiken wie beabsichtigte oder unbeabsichtigte Diskriminierung verhindert werden können. Verzerrungen der Datensätze können die Lernresultate von Algorithmen ändern.

Lösungsansätze bietet hier die EU-Datenschutz-Grundverordnung, die mit ihrem Recht auf Datenauskunft und -korrektur Individuen eine Möglichkeit bietet, Kontrolle auszuüben – allerdings mit zahlreichen Ausnahmen und Einschränkungen, die ergänzende Regulierung notwendig erscheinen lassen. Für die konkrete Umsetzung werden Datenombudsleute oder Datenschutzbevollmächtigte bei öffentlichen und privaten Institutionen empfohlen. Zudem schlagen einige Experten für den globalen Datenhandel Auflagen zu Herkunftsnachweisen vor.

Gesetzgeber, zivilgesellschaftliche Wächter, Forscher: Analyse der Datenbasis institutionalisieren

Der Herkunftsnachweis für Daten führt zu einem weiteren Thema: Die Auszeichnung von Daten hinsichtlich ihrer Beschaffenheit, bekannten Verzerrungen oder Einschränkungen bei der Verwendbarkeit. Zudem haben wir gesehen, dass weitere Verzerrungsmöglichkeiten bei der Verarbeitung von Daten entstehen. Auch die „funktionale Vergleichbarkeit“ von Trainings- und Anwendungsdaten muss gewährleistet werden sowie eine Transparenz über deren Verwendung im Entwicklungs- und Anwendungsprozess.

Die Gewährleistung der Korrektheit, Aktualität und Repräsentativität von Datensets steht am Anfang und bildet einen Raum für Forschung und Innovation. In diesem Bereich könnte durch einen regen Austausch von Best Practices zur Analyse von Datensets ein großer europaweiter Fortschritt erzielt werden.[17]

Gesetzgeber: Gesetzgebung zugunsten von Algorithmenauditing anpassen

Als politische Herausforderung kristallisiert sich hierbei heraus, dass nicht nur der **Zugang zu Algorithmen,** sondern auch der **Zugang zu den Daten** nötig ist, um Auditings durchzuführen. Da sich den Systemen zugrunde liegende Daten im Regelfall in privater Verfügungsgewalt befinden, ist dies oft nicht möglich. Hier gilt es dringend, Transparenzgebote gegenüber staatlichen Institutionen, der Öffentlichkeit oder zivilgesellschaftlicher Wächter- und Verbraucherschutzorganisationen zu prüfen und einzufordern.

Formen des externen Datenzugriffs zum Zwecke des Auditings bergen rechtlichen Reformbedarf: Beispiele aus der Forschung zeigten Möglichkeiten und Grenzen automatisierter Methoden („Webscraping“) oder kollaborativer Forschung („Datenspenden“). Während Erstere schnell an Grenzen der Allgemeinen Geschäftsbedingungen von Systembetreibern stoßen, die unter Umständen zusätzlich durch IT-Sicherheitsgesetz und urheberrechtliche Bestimmungen geschützt sind, mangelt es Letzterem bislang an Repräsentativität und Geschwindigkeit. Manchem Beobachter erscheinen Formen der kollaboraturen Forschung gar als Ausdrucke einer versagenden Rechtsordnung. Hier gilt es**,** rechtliche Restriktionen zu überprüfen und bei Bedarf abzubauen.

Gesetzgeber: Öffentliche Aufsicht über algorithmische Systeme institutionalisieren

Verschiedene vorgestellte Lösungsansätze implizierten die Notwendigkeit der Einrichtung staatlicher Aufsichtsstellen oder zumindest die Definition und Übertragung von Zuständigkeiten an bestehende Institutionen. Im Kern geht es dabei einerseits darum, diejenigen algorithmischen Entscheidungssysteme zu identifizieren, die aus gesamtgesellschaftlicher Perspektive eine Prüfung erfordern. Andererseits müssen dann Verfahren entwickelt werden, um eine solche Prüfung zu ermöglichen. Ob Zulassungsbehörde, Algorithmen-TÜV oder Ratingagentur – Übersicht und Kontrolloptionen müssen in Kooperation mit Systembetreibern hergestellt werden.

Gesetzgeber, institutionelle Nutzer: Widerspruchsverfahren und Verbandsklagerechte schaffen

Eng in Verbindung mit einer Überprüfung algorithmischer Entscheidungssysteme zur Evaluation ihrer gesellschaftlichen Angemessenheit steht die Institutionalisierung von **Widerspruchsverfahren** bzw. **menschlicher Intervention** zum Zwecke der Prüfung der Angemessenheit einer konkreten Entscheidung. Dazu gehört das Recht auf

- Erwirken des Eingreifens einer Person im Sinne der Prüfung einer Entscheidung,
- die Darlegung des eigenen Standpunktes und
- die Anfechtung einer algorithmischen Entscheidung mit gravierenden und/oder rechtlichen Konsequenzen.

Die Verwirklichung dieser Rechte setzt die Erklärbarkeit und Nachvollziehbarkeit algorithmischer Entscheidungsprozesse voraus. Diese können durch eine **Verpflichtung der Systembetreiber zur Dokumentation von Daten und Entscheidungsmethodik** gestärkt werden. Wie bei allen benannten Normen der EU-Datenschutz-Grundverordnung gilt: Zahlreiche Ausnahmen sind vorgesehen, längst nicht alle Arten algorithmischer Entscheidungssysteme abgedeckt. Inwieweit hier weiterer Regulierungsbedarf besteht, ist im Einzelnen zu überprüfen. Dabei sind jenseits formeller Normen auch informelle Anreize und Sanktionen innerhalb der

[17] Telekommunikationsunternehmen und Top Level Domain Registries sind hier vielversprechende Partner, da sie über repräsentative Daten zur Internetnutzung breiter Populationen verfügen. Mit ihrer Hilfe können ggf. Verzerrungen vermieden werden, die auf unvollständigen oder verzerrt gewichteten Datasets beruhen. Um an dieser Stelle Risiken für Privatheit und Datenschutz zu vermeiden, sind Differential-Privacy-Verfahren vielversprechend.

Organisationen bzw. Institutionen zu berücksichtigen, in denen ADM-Systeme Anwendung finden. Darüber hinaus bietet die internationale Debatte wenig Konkretes zur Umsetzung von Widerspruchsverfahren. Gerechtfertigt erscheint in jedem Fall die Prüfung von **Verbandsklagerechten**, damit Wohlfahrtsverbände und Verbraucherschützer Interessen bündeln und fundiert vertreten können.

Gesetzgeber, zivilgesellschaftliche Wächter: Ressourcen zum Algorithmenauditing aufbauen

Um externe Forschung zu stärken, braucht es eine wirksame **Förderung zivilgesellschaftlicher Wächter**. Diese haben sich bislang als wesentlich erwiesen bei der Identifikation neuer Problembereiche und Lösungsansätze. Menschliche wie technische Ressourcen sind notwendig, um die Arbeit zu Sicherheits- und Funktionslücken kontinuierlich zu gewährleisten.

5.3 Die Vielfalt: Diversität algorithmischer Systeme und Prozesse sichern

Die Entwicklung und Anwendung algorithmischer Entscheidungssysteme konzentriert sich bislang im privaten Sektor, in dem die entscheidenden Ressourcen – Daten und Analyseverfahren – konzentriert sind. Damit besteht die Gefahr der Monopolisierung algorithmischer Entscheidungssysteme. Um eine Diversität von Systemen zu ermöglichen und langfristige Entwicklungs- und Kontrollkapazitäten öffentlicher Akteure auszubauen, ist die **freie Verfügbarkeit von Datensätzen** und die **Förderung ihrer Nutzung** von ebenso elementarer Bedeutung wie die **Entwicklung geeigneter Standards**. Sie stellen Möglichkeiten dar, den Informationsasymmetrien zwischen staatlichen und privatwirtschaftlichen Akteuren entgegenzuwirken.

Gesetzgeber: Öffentliche Verfügbarkeit von Datensets fördern

Eine Voraussetzung für die Vielfalt algorithmischer Entscheidungssysteme ist die Zugänglichkeit relevanter Trainingsdatensätze für unterschiedliche Anbieter. Vorschläge dazu umfassen:

- öffentliche Verfügbarkeit von Daten aus öffentlich geförderter Forschung,
- öffentliche Verfügbarkeit von Daten aus dem öffentlichen Sektor, insbesondere aus dem Bereich der öffentlichen Daseinsvorsorge sowie
- öffentliche Verfügbarkeit von Daten aus dem privaten Sektor.

Gesetzgeber, Entwickler, Wissenschaftler: Gemeinwohlorientiertes Datenmanagement institutionalisieren

Bestehende Vorschläge schließen darüber hinausgehende Überlegungen ein, welche

- die Entwicklung von Geschäftsmodellen auf Basis öffentlich verfügbarer Daten,
- die Förderung des Datenmanagements (Datenpflege, Speicherkapazitäten und -anforderungen usw.),
- die Entwicklung von Datenstandards (Metadaten) sowie
- Mindeststandards in der Auszeichnung (Art und Herkunft, Verzerrungen usw.)

betreffen.

Öffentliche Hand: Staatliche Nachfrage nach algorithmischen Systemen zur Vielfaltssicherung nutzen

Die aktiv gestaltende Rolle der öffentlichen Hand im Sinne einer gemeinwohlorientierten Entwicklung algorithmischer Entscheidungssysteme umfasst:

- Entwicklung innovativer algorithmischer Prozesse in der Daseinsvorsorge durch öffentliche Stellen
- Standardentwicklung für algorithmische Entscheidungssysteme im öffentlichen Sektor

- Standardsetzung durch staatliche Nachfrage (Vergabekriterien) – verpflichtende Aufnahme von Standards für Überprüfbarkeit, Folgenabschätzung, offene Standards in öffentlichen Ausschreibungen für algorithmische Systeme
- Förderung von dem Gemeinwohl verpflichteten offenen Technologien

Die Vorschläge machen deutlich, dass dem Staat im Bereich algorithmischer Entscheidungssysteme nicht nur die Rolle eines Regulierers zukommt. Er trägt vielmehr auch als aktiver Gestalter einer positiven Ordnung Verantwortung und sollte dieser – wenn möglich in Kooperation mit anderen Akteuren – unbedingt gerecht werden.

5.4 Der Rahmen: Recht, staatliches Können, individuelle Kompetenzen

Die Übersicht über Herausforderungen und Lösungsansätze im Bereich algorithmischer Entscheidungsfindung macht mehr als deutlich: Sowohl staatliche wie individuelle Kompetenzen müssen grundsätzlich gestärkt werden.

Gesetzgeber: Gesetzlichen Rahmen auf Lücken und effektive Rechtsdurchsetzung prüfen

Generell gilt es, Regelungslücken wie Verbandsklagerechte sowie Lücken bei der Durchsetzung geltenden Rechts zu überprüfen. Dabei gilt es, sowohl die Besonderheiten algorithmischer Entscheidungen wie auch deren Einsatzbereich/Sektor zu berücksichtigen. Dazu zählen auch gesetzliche Verbote und Haftungsfragen. Zwei Kernfragen sind mit Blick auf den rechtlichen Rahmen für algorithmische Systeme zu beantworten:

- Bestehen **Regelungslücken im Recht**?
- Bestehen **Lücken bei der Durchsetzung geltenden Rechts**?

Eine umfassende, systematische Analyse dieser Fragen steht bislang aus.

Die EU-Datenschutz-Grundverordnung stellt erste Anforderungen an die Transparenz algorithmischer Entscheidungen gegenüber Betroffenen im Bereich vollautomatisierter Entscheidungssysteme. Diese stellen im Vergleich zu algorithmischen Assistenzsystemen eher die Ausnahme als die Regel dar. Der Gesetzgeber ist gefordert, weitere Maßnahmen zur Herstellung von Transparenz zu ergreifen und die Ausnahmen von der Verordnung ebenso zu regeln wie solche algorithmischen Entscheidungen, die keine rechtliche Wirkung auf die Betroffenen entfalten.

Zudem gilt es, Regelungen für solche Arten von Daten zu finden, die von der EU-Datenschutz-Grundverordnung nicht gedeckt sind: Zu denken ist hier etwa an Kommunikations- und Bewegungsdaten, die Grundlage neuerer Formen des Profiling sind, sowie die Nutzung anonymisierter Daten.

Gesetzgeber: Staatliches Können stärken

Es gilt einerseits, Übersicht herzustellen über die **Entwicklung, Implementierung und Bewertung** algorithmischer Systeme. Das Unterfangen umfasst mindestens die

- technologische Entwicklung und konkrete Anwendungsfelder
- übergeordnete, systemische Risiken
- Klassifikation von algorithmischen Entscheidungssystemen, beispielsweise in Abhängigkeit ihrer Funktionsweise und Komplexität, dem Einsatzgebiet und ihrer Risiken
- Identifizierung von algorithmischen Entscheidungssystemen, die einer Regulierung im Sinne einer hoheitlichen Aufsicht oder Kontrolle in Form von Auditing, Zertifizierung oder einem Verbot unterliegen sollten
- Aufsicht und Kontrolle über die sachgerechte Verwendung von Daten
- Entwicklung von Sicherheits-, Forschungs- und Anwendungsstandards
- Regelung von Haftbarkeiten und Prüfungsverfahren (Tutt 2016)

Diese Aufgaben könnten beispielsweise von einer **Agentur für algorithmische Systeme** wahrgenommen werden, mit folgenden übergeordneten Zielen: Aufbau und Anwendung rechtlicher, technischer, gesellschaftlicher Kompetenz zu gemeinwohlorientierter Gestaltung algorithmischer Systeme – Beratung, Zertifizierung, Zulassung.

Auf der anderen Seite bedarf es insbesondere im Bereich hoheitlicher und öffentlicher Aufgaben auch des Ausbaus dieser **Kompetenzen bei existierenden Institutionen**. Ziel ist hier einerseits das Verständnis für die Wechselwirkungen zwischen algorithmischer Entscheidungsfindung, Politikprogrammen und gesellschaftlichen Zielen. Andererseits bietet der Kompetenzausbau die Möglichkeit, dass algorithmische Entscheidungssysteme eng mit politischen Prämissen verzahnt werden und spezifische Verwaltungsexpertise mit in das Design algorithmischer Entscheidungsprozesse einfließen kann

Gesetzgeber: Individuelle Sensibilisierung und Kompetenz fördern

Unabhängig vom Bildungshintergrund und dem Vorwissen um algorithmische Prozesse muss jeder Bürger in die Lage versetzt werden, sich gegen fragwürdige Prozesse zu wehren. Individuelle Gestaltungskompetenz muss mit der Stärkung und ggf. dem Aufbau von Institutionen einhergehen, die diese Kompetenz fördern, unterstützen, ergänzen und zum Teil – wo nötig – ausgleichen. Daher ist zum einen der Kenntnisstand der Bevölkerung zu erheben. Andererseits müssen Aus- und Weiterbildungsangebote geschaffen und gefördert werden, zum Beispiel bei Volkshochschulen, Verbraucherschutzorganisationen, Datenschutzbehörden, Stiftungen oder zivilgesellschaftlichen Wächterorganisationen.

5.5 Jetzt handeln!

Die Vorschläge konzentrieren sich zumeist auf technische und rechtliche Aspekte, ergeben in der Zusammenschau jedoch recht integrative Lösungsansätze, in denen staatlichen, wirtschaftlichen und zivilgesellschaftlichen Akteuren unterschiedliche **Verantwortung** zukommt. Vielen konstruktiven Vorschlägen mangelt es noch an der **Konkretisierung** und einer konkreten Aufgabenverteilung bzw. **Institutionalisierung**. Das parallele Erproben unterschiedlicher Lösungsansätze scheint dabei so Erfolg versprechend wie eine Koordination der Aktivitäten.

Der Überprüfung vorhandener Lösungsansätze, inklusive rechtlicher Restriktionen, sollte im Kontext der Problemanalyse Vorzug gegenüber voreiliger Regulierung gegeben werden. Insgesamt gilt es bei der Erwägung von Regulierung, auch die Unzulänglichkeiten heutiger Entscheidungssysteme zu debattieren und auf mögliche **Chancen algorithmischer Entscheidungen** zu fokussieren.

Zuletzt gilt es zu betonen: Die Übersicht über Handlungsbedarf und Lösungsansätze im Kontext algorithmischer Entscheidungsverfahren hat gezeigt: **Vielfältige Maßnahmen und Methoden bieten Möglichkeiten gesellschaftlicher Mitgestaltung, Intervention und Kontrolle**. Keineswegs scheint der Mensch der Maschine ausgeliefert. Allerdings gilt es nun, Chancen und Risiken im Einzelfall, d. h. unter Berücksichtigung des Anwendungsbereichs, der Komplexität und der Autonomie eines Systems, zu überprüfen und ggf. spezifische Handlungsoptionen zu entwickeln und zu erproben.

6 Literatur

§ 34 BDSG – Einzelnorm (o. J.). https://www.gesetze-im-internet.de/bdsg_1990/__34.html (Download 29.3.2018).

American Civil Liberties Union (ACLU) (2017). „Sandvig v. Sessions – Challenge to CFAA Prohibition on Uncovering Racial Discrimination Online“. https://www.aclu.org/cases/sandvig-v-sessions-challenge-cfaa-prohibition-uncovering-racial-discrimination-online (Download 8.1.2018).

AI Now Institute (2017). „AI Now Public Symposium“. https://www.youtube.com/watch?v=ORHe3dMvR2c (Download 22.4.2018).

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman und Dan Mané (o. J.). „Concrete Problems in AI Safety“. https://arxiv.org/pdf/1606.06565.pdf (Download 22.4.2018).

Angwin, Julia, Jeff Larson, Surya Mattu und Lauren Kirchner (2016). „Machine Bias: There’s Software Used Across the Country to Predict Future Criminals. And it’s Biased Against Blacks“. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (Download 11.12.2016).

Association for Computing Machinery (ACM) (2017). „Statement on Algorithmic Transparency and Accountability“. https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf (Download 22.4.2018).

Bakshy, Eytan, Solomon Messing und Lada A. Adamic (2015). „Exposure to ideologically diverse news and opinion on Facebook“. *Science*, *348*(6239), 1130–1132. https://doi.org/10.1126/science.aaa1160 (Download 22.4.2018).

Barocas, Solon, und Andrew D. Selbst (2014). „Big Data’ s Disparate Impact“. *California Law Review* (104) 3. 1–57. https://doi.org/10.15779/Z38BG31 (Download 22.4.2018).

Beirat Integration (2013). „‚Soziale Teilhabe‘ Handlungsempfehlungen des Beirats der Integrationsbeauftragten“. Die Beauftragte der Bundesregierung für Migration, Flüchtlinge und Integration. http://www.bagiv.de/pdf/soziale-teilhabe-empfehlungen-beirat.pdf (Download 22.4.2018).

Berghahn, Sabine, Vera Egenberger, Micha Klapp, Alexander Klose, Doris Liebscher, Linda Supik und Alexander Tischbirek (2016). *Evaluation des Gleichbehandlungsgesetzes.* Hrsg. Antidiskriminierungsstelle des Bundes. Berlin. (Auch online unter https://www.antidiskriminierungsstelle.de/SharedDocs/Downloads/DE/publikationen/AGG/AGG_Evaluation.pdf?__blob=publicationFile&v=15, Download 22.4.2018.)

Bertelsmann Stiftung (2011). *Soziale Gerechtigkeit in der OECD – Wo steht Deutschland? Sustainable Governance Indicators 2011*. Gütersloh. (Auch online unter http://news.sgi-network.org/uploads/tx_amsgistudies/SGI11_Social_Justice_DE.pdf, Download 22.4.2018.)

Bertelmann Stiftung (2017). „Förderung – Mehr Transparenz und zivilgesellschaftliche Kontrolle von Algorithmen“. https://www.bertelsmann-stiftung.de/de/unsere-projekte/teilhabe-in-einer-digitalisierten-welt/projektnachrichten/foerderung-von-algorithmwatch/ (Download 12.4.2018).

Beuth, Patrick (2017). „Bombenbauer, die Aceton gekauft haben, kauften auch ... “ *Zeit Online* 4.11. http://www.zeit.de/digital/internet/2017-11/amazon-terrorverdaechtiger-sprengstoff-zutaten-bestellt/komplettansicht?print (Download 22.4.2018).

Böttcher, Björn, Daniel Klemm und Carlo Velten (2017). *Machine Learning im Unternehmenseinsatz*. Hrsg. Crisp Research AG. Kassel. (Auch online unter https://www.unbelievable-machine.com/downloads/studie-machine-learning.pdf, Download 10.5.2018.)

Brennan Center for Justice (2017). „Brennan Center for Justice v. New York Police Department“. https://www.brennancenter.org/legal-work/brennan-center-justice-v-new-york-police-department (Download 4.1.2018).

Briseno, Cinthia (2018). „Wie Algorithmen Menschen vor einem frühzeitigen Tod bewahren können“. *Algorithmenethik* 15.3. https://algorithmenethik.de/2018/03/15/wie-algorithmen-menschen-vor-einem-fruehzeitigen-tod-bewahren-koennen/ (Download 13.4.2018).

Buermeyer, Ulf (2016). „‚Digitaler Hausfriedensbruch‘: IT-Straf-recht auf Abwegen“. *Legal Tribune Online* 6.10. https://www.lto.de/recht/hintergruende/h/entwurf-straftatbestand-digitaler-hausfriedensbruch-botnetze-internet/ (Download 22.4.2018).

Bundesamt für Migration und Flüchtlinge (2017). „Moderne Technik in Asylverfahren“. 26.7. https://www.bamf.de/SharedDocs/Meldungen/DE/2017/20170726-am-vorstellung-modellprojekt-bamberg.html (Download 18.1.2018).

Bundesärztekammer (2015). „Musterberufsordnung für die in Deutschland tätigen Ärztinnen und Ärzte“. http://www.bundesaerztekammer.de/fileadmin/user_upload/downloads/pdf-Ordner/MBO/MBO_02.07.2015.pdf (Download 22.4.2018).

Bundesministerium für Wirtschaft und Energie (2016). *Digitale Strategie 2025*. Berlin. (Auch online unter http://www.bmwi.de/Redaktion/DE/Publikationen/Digitale-Welt/digitale-strategie-2025.pdf, Download 22.4.2018.)

Bundesregierung (2017). „Strassenverkehrsgesetz – Automatisiertes Fahren auf dem Weg“. https://www.bundesregierung.de/Content/DE/Artikel/2017/01/2017-01-25-automatisiertes-fahren.html (Download 16.1.2018).

Calo, Ryan (2017). „Artificial Intelligence Policy: A Roadmap“. *SSRN Scholarly Paper* No. ID 3015350). Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=3015350 (Download 22.4.2018).

Campolo, Alex, Madelyn Sanfilippo, Meredith Whittaker und Kate Crawford (2017). „AI Now 2017 Report (No. 2)“. New York NY: AI Now Institute. https://ainowinstitute.org/AI_Now_2017_Report.pdf (Download 22.4.2018).

Caplan, Robyn, Laura Reed und Alexandra Mateescu (2016). „Who Controls the Public Sphere in an Era of Algorithms – Workshop Summary“. Gehalten auf der Who Controls the Public Sphere in an Era of Algorithms, Data & Society Research Institute. https://datasociety.net/pubs/ap/WorkshopNotes_PublicSphere_2016.pdf (Download 22.4.2018).

Cave, Stephen (2017). „Written evidence – Leverhulme Centre for the Future of Intelligence“. http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/artificial-intelligence-committee/artificial-intelligence/written/69702.html (Download 22.4.2018).

Christl, Wolfie (2014). *Kommerzielle digitale Überwachung im Alltag. Erfassung, Verknüpfung und Verwertung persönlicher Daten im Zeitalter von Big Data: Internationale Trends, Risiken und Herausforderungen anhand ausgewählter Problemfelder und Beispiele*. Wien: Cracked Labs. (Auch online unter http://crackedlabs.org/dl/Studie_Digitale_Ueberwachung.pdf, Download 22.4.2018.)

Christl, Wolfie (2017). *How Companies Use Personal Data Against People – Automated disadvantage and personalized manipulation?* Working Paper by Cracked Labs. Wien. (Auch online unter http://crackedlabs.org/dl/CrackedLabs_Christl_DataAgainstPeople.pdf, Download 22.4.2018.)

Citron, Danielle Keats (2008). „Technological Due Process“. *Washington University Law Review* (85) 1. 1249.

Citron, Danielle Keats, und Frank Pasquale (2014). „The Scored Society: Due Process for Automated Predictions“. *Washington Law Review*, (89) 1. 1–33.

City of Chicago (2017). „Strategic Subject List“. https://data.cityofchicago.org/Public-Safety/Strategic-Subject-List/4aki-r3np (Download 22.4.2018).

Cohen, I. Glenn, Ruben Amarasingham, Anand Shah, Bin Xie und Bernard Lo (2014). „The Legal And Ethical Concerns That Arise From Using Complex Predictive Analytics In Health Care“. *Health Affairs* (*33)* 7. 1139–1147. https://doi.org/10.1377/hlthaff.2014.0048 (Download 22.4.2018).

Council of Europe – Committee of experts on internet intermediaries (2017). „Study on the Human Rights Dimensions of Automated Data Processing Techniques (in Particular Algorithms) and Possible Regulatory

Implications“. https://rm.coe.int/study-hr-dimension-of-automated-data-processing-incl-algorithms/168075b94a (Download 22.4.2018).

Dahllof, Staffan, Orr Hirschauge, Hagar Shezaf, Jennifer Baker und Nikolaj Nielsen (2017). „EU states copy Israel’s , predictive policing‘“. *EUobserver* 6.10. https://euobserver.com/justice/139277 (Download 22.4.2018).

Data & Society. Data Society Research Institute – Written evidence (AIC0221), Pub. L. No. AIC0221 und Select Committee on Artificial Intelligence (2017). „Response to UK House of Lords Call for Evidence“. http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/artificial-intelligence-committee/artificial-intelligence/written/70517.html (Download22.4.2018).

Deng, Jia, Wei Dong, Richard Socher, Li-Ja Li, Kai Li und Li Fei-Fei (2009). „ImageNet: A Large-Scale Hierarchical Image Database“. *Computer Vision and Pattern Recognition 2009.* CVPR 2009. IEEE Conference on. 248–255. http://www.image-net.org/papers/imagenet_cvpr09.pdf (Download 22.4.2018).

Deutscher Bundestag 19. Wahlperiode (2018). „Schriftliche Fragen mit den in der Woche vom 29. Januar 2018 eingegangenen Antworten der Bundesregierung (No. Drucksache 19/605)“.

Dewes, Andreas (2018). „Begutachtung des Arbeitspapiers ‚Damit Maschinen den Menschen dienen‘“. (unveröffentlicht).

Diakopoulos, Nicholas (2016). „Accountability in Algorithmic Decision Making“. *Commun. ACM* (59) 2. 56–62. https://doi.org/10.1145/2844110 (Download 22.4.2018).

Dickey, Megan Rose (2016). „Police are increasingly using social media surveillance tools“. https://techcrunch.com/2016/09/23/police-are-increasingly-using-social-media-surveillance-tools/ (Download 22.4.2018).

Diedrich, Oliver (2014). „Linux Foundation finanziert OpenSSL-Entwickler“. https://www.heise.de/ho/meldung/Linux-Foundation-finanziert-OpenSSL-Entwickler-2213936.html (Download 10.11.2017).

Doctorow, Cory (2018). „Two years later, Google solves ‚racist algorithm‘problem by purging ‚gorilla‘ label from image classifier“. *boingboing* 11.1. https://boingboing.net/2018/01/11/gorilla-chimp-monkey-unpersone.html (Download 22.4.2018).

Dräger, Jörg, und Ralph Müller-Eiselt (2015). *Die digitale Bildungsrevolution. Der radikale Wandel des Lernens und wie wir ihn gestalten können*. München.

Dreyer, Stephan, und Wolfgang Schulz (2018). *Was bringt die Datenschutz-Grundverordnung für automatisierte Entscheidungssysteme?* Impuls Algorithmenethik No. #5. Hrsg. Bertelsmann Stiftung. Gütersloh.

Eckersley, Peter, Jeremy Gillula und Jamie Williams (o. J.). „Written evidence – Electronic Frontier Foundation“ “. http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/artificial-intelligence-committee/artificial-intelligence/written/69720.html (Download 22.4.2018).

Europäisches Parlament, und Rat der Europäischen Union (2016). „Verordnung (EU) 2016/679 des Europäischen Parlaments und des Rates vom 27. April 2016 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten, zum freien Datenverkehr und zur Aufhebung der Richtlinie 95/46/EG (Datenschutz-Grundverordnung), Pub. L. No. Verordnung (EU) 2016/679 (2016)“. http://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX:32016R0679 (Download 22.4.2018).

Executive Office of the President, President’s Council of Advisors on und Science and Technology (2014). „Big Data and Privacy: A technological Perspective“. https://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf (Download 22.4.2018).

Executive Office of the President, National Science and Technology Council und Committee on Technology (2016). „Preparing for the Future of Artificial Intelligence“. https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf (Download 22.4.2018).

Fairness, Accountability, and Transparency in Machine Learning (FAT/ML) (2016). „Principles for Accountable Algorithms and a Social Impact Statement for Algorithms“. http://www.fatml.org/resources/principles-for-accountable-algorithms (Download 1.2.2017).

Ferguson, Andrew G. (2017). „Policing Predictive Policing“. *Washington University Law Review* (94) 5. 1113–1195.

Frick SJ, Eckhard (2018). „Welche Philosophie braucht die Medizin?“ *Stimmen der Zeit* 2. 100–109.

forum (2017). „Autonomes Fahren ist nach Änderung des Straßenverkehrsgesetzes (StVG) erlaubt“. https://www.forum-verlag.com/themenwelten/kommunales/autonomes-fahren-ist-nach-aenderung-des-strassenverkehrsgesetzes-stvg-erlaubt (Download 16.1.2018).

frankro (2010). „Datenbrief“.

Free Credit Scores Estimator from myFICO (o. J.). http://www.myfico.com/fico-credit-score-range-estimator/ (Download 29.3.2018).

Future of Life Institute (2017). „Asilomar AI Principles“. https://futureoflife.org/ai-principles/ (Download 5.2.2017).

Future of Privacy Forum (2017). „Unfairness by Algorithm: Distilling the Harms of Automated Decision-Making“. Gehalten auf der RightsCon, Brüssel. https://fpf.org/2017/12/11/unfairness-by-algorithm-distilling-the-harms-of-automated-decision-making/ (Download 22.4.2018).

Gallwitz, Florian (2017a). „Eine Polemik: Wie man mit einem würfelnden Schimpansen Terroristen fängt". *Algorithmenethik* 21.12. https://algorithmenethik.de/2017/12/21/eine-polemik-wie-man-mit-einem-wuerfelnden-schimpansen-terroristen-faengt/ (Download 1.2.2018).

Gallwitz, Florian (2017b). „Begutachtung des Arbeitspapiers ‚Damit Maschinen den Menschen dienen'" (unveröffentlicht).

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III und Kate Crawford (2018). „Datasheets for Datasets". http://jamiemorgenstern.com/papers/datasheet.pdf (Download 22.4.2018).

Georgieva, Petia (2017). „Written evidence – IEEE European Public Policy Initiative – Working Group on ICT". http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/artificial-intelligence-committee/artificial-intelligence/written/69590.html (Download 22.4.2018).

Gershgorn, Dave (2017). „The data that transformed AI research – and possibly the world". https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world/ (Download 5.11.2017).

Goodman, Bryce W. (2015). „A Step Towards Accountable Algorithms?: Algorithmic Discrimination and the European Union General Data Protection". Gehalten auf der 29th Conference on Neural Information Processing Systems, Montreal (Kanada). http://www.mlandthelaw.org/papers/goodman1.pdf (Download 22.4.2018).

Gunning, David (2016). „Explainable Artificial Intelligence (XAI)". https://www.cc.gatech.edu/~alanwags/DLAI2016/%28Gunning%29%20IJCAI-16%20DLAI%20WS.pdf (Download 22.4.2018).

Harris, Elizabeth A. (2016). „Court Vacates Long Island Teacher's Evaluation Tied to Test Scores". *The New York Times* 10.5. https://www.nytimes.com/2016/05/11/nyregion/court-vacates-long-island-teachers-evaluation-tied-to-student-test-scores.html (Download 22.4.2018).

Harvey, Hugh (2018). „Why AI will not replace radiologists". https://towardsdatascience.com/why-ai-will-not-replace-radiologists-c7736f2c7d80 (Download 7.3.2018).

Heaton, Brian (2015). „New York City Fights Fire with Data". Government Technology 15.5. http://www.govtech.com/public-safety/New-York-City-Fights-Fire-with-Data.html (Download 22.4.2018).

Bundesgesetzblatt (1996). Internationaler Pakt über wirtschaftliche, soziale und kulturelle Rechte. (1966, Dezember 19). http://www.institut-fuer-menschenrechte.de/fileadmin/user_upload/PDF-Dateien/Pakte_Konventionen/ICESCR/icescr_de.pdf (Download 22.4.2018)

Jaume-Palasí, Lorena (2017). „Diskriminierung hängt nicht vom Medium ab“. *AlgorithmWatch* 3.7. https://algorithmwatch.org/de/diskriminierung-haengt-nicht-vom-medium-ab/ (Download 6.2.2018).

Jaume-Palasí, Lorena, und Matthias Spielkamp. (2017). „Ethics and algorithmic processes for decision making and decision support“. *AlgorithmWatch*. https://algorithmwatch.org/wp-content/uploads/2017/06/AlgorithmWatch_Working-Paper_No_2_Ethics_ADM.pdf (Download 10.5.2018).

Jedrzej, Niklas, Karolina Sztandar-Sztanderska und Katarzyna Szymielewicz (2015). *Profiling the Unemployed in Poland: Social and Political Implications of Algorithmic Decision Making*. Warschau: Fundacja Panoptykon. (Auch online unter https://panoptykon.org/sites/default/files/leadimage-biblioteka/panoptykon_profiling_report_final.pdf, Download 22.4.2018.)

juris (2018). „Schutz vor digitalem Hausfriedensbruch“. https://www.juris.de/jportal/portal/t/1l1i/page/homerl.psml?nid=jnachr-JUNA180300598&cmsuri=%2Fjuris%2Fde%2Fnachrichten%2Fzeigenachricht.jsp (Download 22.4.2018).

Kahneman, Daniel, Andrew M. Rosenfield, Linnea Gandhi und Tom Blaser (2016). „Noise: How to Overcome the High, Hidden Cost of Inconsistent Decision Making“. https://hbr.org/2016/10/noise (Download 25.3.2018).

Kirchner, Lauren (2017). „Putting Crime Scene DNA Analysis on Trial. *ProPoblica* 11.11. https://www.propublica.org/article/putting-crime-scene-dna-analysis-on-trial (Download 1.4.2018).

Kitchin, Rob (2016). „Thinking critically about and researching algorithms“. *Information, Communication & Society* (20) 1. 1–16.

Krafft, Tobias D., Michael Gamer, Marcel Laessing und Katharina Anna Zweig (2017). „1. Zwischenbericht Datenspende: Filterblase geplatzt? Kaum Raum für Personalisierung bei Google-Suchen zur Bundestagswahl 2017“. *AlgorithmWatch*. https://algorithmwatch.org/wp-content/uploads/2017/09/1_Zwischenbericht__final.pdf (Download 22.4.2018).

Kroll, Joshua A., Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson und Harlan Yu (2017). „Accountable Algorithms“. *University of Pennsylvania Law Review* (165) 3. 633–705.

Kunichoff, Yana, und Patrick Sier (2017). „The Contradictions of Chicago Police’s Secretive List“. http://www.chicagomag.com/city-life/August-2017/Chicago-Police-Strategic-Subject-List/ (Download 22.4.2018).

Laskowski, Nicole (2017). „Machine learning’s training data is a security vulnerability“. *TechTarget* 31.10. http://searchcio.techtarget.com/news/450429272/Machine-learnings-training-data-is-a-security-vulnerability (Download 22.4.2018).

Lazer, David (2015). „The rise of the social algorithm“. *Science* (348) 6239. https://doi.org/10.1126/science.aab1422 (Download 22.4.2018).

Lenk, Klaus (2016). „Die neuen Instrumente der weltweiten digitalen Governance“. *Verwaltung und Management* (*22)* 5. 227–240.

Lischka, Konrad, und Anita Klingel (2017). *Wenn Maschinen Menschen bewerten.* Impuls Algorithmenethik No. #1. Hrsg. Bertelsmann Stiftung. Gütersloh. https://www.bertelsmann-stiftung.de/de/publikationen/publikation/did/wenn-maschinen-menschen-bewerten/ (Download 22.4.2018).

Lischka, Konrad, und Christian Stöcker (2017). *Digitale Öffentlichkeit. Wie algorithmische Prozesse den gesellschaftlichen Diskurs beeinflussen.* Impuls Algorithmenethik No. #3. Hrsg. Bertelsmann Stiftung. Gütersloh. https://doi.org/10.11586/2017028 (Download 22.4.2018).

Martini, Mario (2017). „Algorithmen als Herausforderung für die Rechtsordnung“. *JuristenZeitung* (72) 21. 1017–1026.

Mateescu, Alexandr, Douglas Brunton, Alex Rosenblat, Desmond Patton, Zachary Gold und Danah Boyd (2015). „Social Media Surveillance and Law Enforcement“. New York NY: Data & Society Research Institute. http://www.datacivilrights.org/pubs/2015-1027/Social_Media_Surveillance_and_Law_Enforcement.pdf (Download 22.4 2018).

Meyer, Thomas (2016). „Gleichheit – warum, von was und wie viel?“ *Neue Gesellschaft/Frankfurter Hefte* 11. 42–46.

Mittelstadt, Brent (2016). „Auditing for Transparency in Content Personalization Systems“. *International Journal of Communication* (10) 0. 4991–5002.

Mittelstadt, Brent, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter und Luciano Floridi (2016). „The Ethics of Algorithms: Mapping the Debate“. http://philpapers.org/archive/MITTEO-12.pdf (Download 10.5.2018).

Moravec, H. (1998). When will computer hardware match the human brain. Journal of evolution and technology, 1(1), 10.

Mulgan, Geoff (2016). „A machine intelligence commission for the UK: how to grow informed public trust and maximise the positive impact of smart machines“. https://www.nesta.org.uk/sites/default/files/a_machine_intelligence_commission_for_the_uk_-_geoff_mulgan.pdf (Download 22.4.2018).

National Science and Technology Council (2016). „The National Artificial Intelligence Research and Development Strategic Plan“. https://www.nitrd.gov/PUBS/national_ai_rd_strategic_plan.pdf (Download 22.4.2018).

New York City Independent Budget Office (2016). „A Look at New York City’s Public High School Choice Process“. http://www.ibo.nyc.ny.us/iboreports/preferences-and-outcomes-a-look-at-new-york-citys-public-high-school-choice-process.pdf (Download 22.4.2018).

O'Neil, Cathy (2016). *Weapons of math destruction: how big data increases inequality and threatens democracy* New York NY: Crown.

O'Neil, Cathy (2017). „Don't Grade Teachers With a Bad Algorithm". *Bloomberg.com* 15.5. https://www.bloomberg.com/view/articles/2017-05-15/don-t-grade-teachers-with-a-bad-algorithm (Download 22.4.2018).

Open Knowledge Foundation Deutschland (o. J.). „Prototype Fund". https://okfn.de/projekte/prototypefund/ (Download 3.11.2017).

Otto, Philipp (2017). „Leben im Datenraum – Handlungsaufruf für eine gesellschaftlich sinnvolle Nutzung von Big Data". *Perspektiven der digitalen Lebenswelt*. Hrsg. Herrmann Hill, Dieter Kugelmann und Mario Martini. Baden-Baden. 9–36. (Auch online unter https://irights-lab.de/wp-content/uploads/2017/06/Leben-im-Datenraum_Philipp-Otto_Perspektiven-der-digitalen-Lebenswelt_HillMartiniKugelmann.pdf, Download 22.4.2018).

Otto, Philipp (2018). „Begutachtung des Arbeitspapiers 'Damit Maschinen den Menschen dienen'" (unveröffentlicht).

Pasquale, Frank (2010). „Beyond innovation and competition: The need for qualified transparency in internet intermediaries". *Northwestern University Law Review* (104) 105. 105–171.

Pasquale, Frank (2016). *The Black Box Society: The Secret Algorithms That Control Money and information* (Reprint). Cambridge MA und London, England: Harvard University Press.

Passig, Kathrin (2017). „Fünfzig Jahre Black Box". *Merkur* (71) 823. 16–30.

Powles, Julia (2017). „New York City's Bold, Flawed Attempt to Make Algorithms Accountable". *The New Yorker* 21.12. https://www.newyorker.com/tech/elements/new-york-citys-bold-flawed-attempt-to-make-algorithms-accountable (Download 22.4.2018).

Prainsack, Barbara (2017). „Research for Personalized Medicine: Time for Solidarity". *Medicine and Law. World Association for Medical Law* (36) 1. 87–98.

Ramge, Thomas (2018). *Mensch und Maschine. Wie Künstliche Intelligenz und Roboter unser Leben verändern*. Stuttgart.

Rohde, Noëlle (2017). „In Australien prüft eine Software die Sozialbezüge – und erfindet Schulden für 20.000 Menschen". *Algorithmenethik* 25.10. https://algorithmenethik.de/2017/10/25/in-australien-prueft-eine-software-die-sozialbezuege-und-erfindet-schulden-fuer-20-000-menschen/ (Download 1.11.2017).

Roßnagel, Alexander (2017). „Zusätzlicher Arbeitsaufwand für die Aufsichtsbehörden der Länder durch die Datenschutz-Grundverordnung". https://www.datenschutzzentrum.de/uploads/dsgvo/2017-Rossnagel-Gutachten-Aufwand-Datenschutzbehoerden.pdf (Download 22.4.2018).

Royal Society (2017). *Machine learning: the power and promise of computers that learn by example*. London: The Royal Society. (Auch online unter https://royalsociety.org/~/media/policy/projects/machine-learning/publications/machine-learning-report.pdf, Download 22.4.2018.)

Russel, Stuart, Daniel Dewey und Max Tegmark (2015). „Research Priorities for Robust and Beneficial Artificial Intelligence. Association for the Advancement of Artificial Intelligence". https://futureoflife.org/data/documents/research_priorities.pdf?x56934 (Download 22.4.2018).

Russel, Stuart, und Peter Norvig (2012). *Künstliche Intelligenz. Ein moderner Ansatz.* 3., aktualisierte Auflage. München.

Sandvig, Christian (2015). „The Facebook ‚It's Not Our Fault' Study". *Social Media Collective Research Blog* 7.5. https://socialmediacollective.org/2015/05/07/the-facebook-its-not-our-fault-study/ (Download 20.1.2017).

Sandvig, Ckristian, Kevin Hamilton, Karrie Karahalios und Cendric Langbort (2014). „Auditing algorithms: Research methods for detecting discrimination on internet platforms". *Data and Discrimination: Converting Critical Concerns into Productive Inquiry*. https://pdfs.semanticscholar.org/b722/7cbd34766655dea10d0437ab10df3a127396.pdf (Download 22.4.2018).

Scherer, Matthew U. (2016). „Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies". *Harvard Journal of Law & Technology* (29) 2. 354–400.

Scherer, Matthew U. (2017). „Public Risk Management for A.I.: The Path Forward". Gehalten auf der Beneficial AI 2017 – Asilomar Conference 2017. https://futureoflife.org/wp-content/uploads/2017/01/Matthew-Scherer.pdf?x56934 (Download 22.4.2018).

Schneider, Jan, Ruta Yemane und Martin Weinmann (2014). *Diskriminierung am Ausbildungsmarkt: Ausmaß, Ursachen und Handlungsperspektiven*. Saarbrücken. (Auch online unter http://www.svr-migration.de/wp-content/uploads/2014/11/SVR-FB_Diskriminierung-am-Ausbildungsmarkt.pdf, Download 22.4.2018.)

Schuetze, Julia (2018). *Warum dem Staat IT-Sicherheitsexpert:innen fehlen. Eine Analyse des IT-Sicherheitskräftemangels im Öffentlichen Dienst*. Hrsg. Stiftung Neue Verantwortung. Berlin. (Auch online unter https://www.stiftung-nv.de/sites/default/files/it-sicherheitsfachkraeftemangel.pdf, Download 22.4.2018.)

Selbst, Andrew D. (2016). *Disparate Impact in Big Data Policing* (SSRN Scholarly Paper No. ID 2819182). Rochester NY: Social Science Research Network. (Auch online unter http://papers.ssrn.com/abstract=2819182, Download 22.4.2018.)

Shneiderman, Ben (2016). „Opinion: The dangers of faulty, biased, or malicious algorithms requires independent oversight". *Proceedings of the National Academy of Sciences* (113) 48. 13538–13540. https://doi.org/10.1073/pnas.1618211113 (Download 22.4.2018).

Singer, Natasha (2015). „Bringing Big Data to the Fight Against Benefits Fraud“. *The New York Times* 20.12. https://www.nytimes.com/2015/02/22/technology/bringing-big-data-to-the-fight-against-benefits-fraud.html (Download 22.4.2018).

Spindler, Gerald (2015). „Stellungnahme zum Gesetz zur Verbesserung der zivilrechtlichen Durchsetzung von verbraucherschützenden Vorschriften des Datenschutzrechts – RegE BT-Drucks. 18/4631“. http://webarchiv.bundestag.de/cgi/show.php?fileToLoad=4246&id=1269 (Download 22.4.2018).

Stalder, Felix (2017). „Algorithmen, die wir brauchen“. *Netzpolirik.org* 15.1. 16. Januar 2017. https://netzpolitik.org/2017/algorithmen-die-wir-brauchen/ (Download 22.4.2018).

Staltz, André (2017). „The Web began dying in 2014, here’s how. https://staltz.com/the-web-began-dying-in-2014-heres-how.html (Download 5.11.2017).

Stone, Peter, Rodney Brooks, Erik Brynjolfsson, Ryan Calo, Oren Etzioni, Greg Hager, Julia Hirschberg, Shivaram Kalyanakrishan, Ece Kamar, Sarit Kraus, Kevin Leyton-Brown, David Parkes, William Press, AnnaLee Saxenian, Julie Shah, Milind Tambe und Astro Teller (2016). „Artificial Intelligence and Life in 2030“. Stanford CA: Stanford University. https://ai100.stanford.edu/sites/default/files/ai_100_report_0831fnl.pdf (Download 22.4.2018).

Tene, Omer, und Jules Polonetsky (2017). „Taming the Golem: Challenges of Ethical Algorithmic Decision-Making“. *North Carolina Journal of Law & Technology* (19) 1. 125–173.

The New York City Council (2018). „A Local Law in relation to automated decision systems used by agencies, Pub. L. No. Int 1686-2017 (2018)“. http://legistar.council.nyc.gov/LegislationDetail.aspx?ID=3137815&GUID=437A6A6D-62E1-47E2-9C42-461253F9C6D0 (Download 22.4.2018).

Torres, Phil (2017). „The Divide Between People Who Hate and Love Artificial Intelligence Is Not Real“. *Motherboard* 27.10. https://motherboard.vice.com/en_us/article/7x48kg/the-divide-between-people-who-hate-and-love-artificial-intelligence-is-not-real (Download 2.11.2017).

Tullis, Tracy (2014). „How Game Theory Helped Improve New York City’s High School Application Process“. *The New York Times* 5.12. https://www.nytimes.com/2014/12/07/nyregion/how-game-theory-helped-improve-new-york-city-high-school-application-process.html (Download 22.4.2018).

Tutt, Andrew (2016). „An FDA for Algorithms“ (SSRN Scholarly Paper No. ID 2747994). Rochester NY: Social Science Research Network. https://papers.ssrn.com/abstract=2747994 (Download 22.4.2018).

University of Utah Honors (2017). „Justice.exe“. http://justiceexe.com/ (Download 22.4.2018).

Vieth, Kilian, und Ben Wagner (2017). *Teilhabe, ausgerechnet.* Impuls Algorithmenethik No. #2. Hrsg. Bertelsmann Stiftung). Gütersloh. (Auch online unter https://doi.org/10.11586/2017027, Download 22.4.2018).

Wachter, Sandra, Brent Mittelstadt und Chris Russell (2017). „Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR“ (SSRN Scholarly Paper No. ID 3063289). Rochester NY: Social Science Research Network. https://papers.ssrn.com/abstract=3063289 (Download 10.5.2018).

Web Foundation (2017). „Algorithmic Accountability. Applying the concept to different country contexts“. Washington DC: World Wide Web Foundation. http://webfoundation.org/docs/2017/07/Algorithms_Report_WF.pdf (Download 22.4.2018).

Williams, Jamie (2017). „EFF to Court: Accessing Publicly Available Information on the Internet Is Not a Crime“. *Electronic Frontier Foundation* 11.12. https://www.eff.org/deeplinks/2017/12/eff-court-accessing-publicly-available-information-internet-not-crime (Download 9.1.2018).

Williams, Jamie (2018). „Ninth Circuit Doubles Down: Violating a Website’s Terms of Service Is Not a Crime“. *Electronic Frontier Foundation* 10.1. https://www.eff.org/deeplinks/2018/01/ninth-circuit-doubles-down-violating-websites-terms-service-not-crime (Download 11.1.2018).

Zweig, Katharina Anna (2016). „2. Arbeitspapier: Überprüfbarkeit von Algorithmen“. *AlgorithmWatch* 7.7. http://algorithmwatch.org/zweites-arbeitspapier-ueberpruefbarkeit-algorithmen/ (Download 7.9.2016).

Zweig, Katharina Anna (2018). *Wo Maschinen irren können – Fehlerquellen und Verantwortlichkeiten in Prozessen algorithmischer Entscheidungsfindung*. Hrsg. Bertelsmann Stiftung. Gütersloh. (Auch online unter https://doi.org/10.11586/2018006, Download 10.5.2018.)

7 Über die Autoren

Julia Krüger: Frei arbeitende Sozialwissenschaftlerin aus Berlin. Sie interessiert sich für Internet- und Digitalisierungspolitik im internationalen Vergleich, insbesondere für die Themen: Inhalte Regulierung und Manipulation, Daten- und Verbraucherschutz sowie Algorithmen und maschinelles Lernen.

Konrad Lischka schreibt seit 1999 über die digitale Gesellschaft – Bücher, Essays und Blogs. Nach dem Diplomstudium der Journalistik und der Ausbildung an der Deutschen Journalistenschule arbeitete er als Chefredakteur des bücher Magazins und stellvertretender Ressortleiter Netzwelt bei Spiegel Online. Danach Wechsel in die Medien- und Netzpolitik als Referent Digitale Gesellschaft in der Staatskanzlei Nordrhein-Westfalen, seit 2016 Projektmanager bei der Bertelsmann Stiftung, Projektleiter Ethik der Algorithmen.

8 Impulse Algorithmenethik

Alle Veröffentlichungen sind abrufbar unter: https://algorithmenethik.de/impulse/

Impuls Algorithmenethik #1: Lischka, Konrad, und Anita, Klingel (2017). *Wenn Maschinen Menschen bewerten*. Hrsg. Bertelsmann Stiftung. Gütersloh. (Auch online unter https://doi.org/10.11586/2017025, Download 10.5.2018.)

Impuls Algorithmenethik #2: Vieth, Kilian, und Ben Wagner (2017). *Teilhabe, ausgerechnet*. Hrsg. Bertelsmann Stiftung. Gütersloh. (Auch online unter https://doi.org/10.11586/2017027, Download 10.5.2018.)

Impuls Algorithmenethik #3: Lischka, Konrad, und Christian Stöcker (2017). *Digitale Öffentlichkeit*. Hrsg. Bertelsmann Stiftung. Gütersloh. (Auch online unter https://doi.org/10.11586/2017028, Download 10.5.2018.)

Impuls Algorithmenethik #4: Zweig, Katharina Anna (2018). *Wo Maschinen irren können. Fehlerquellen und Verantwortlichkeiten in Prozessen algorithmischer Entscheidungsfindung*. Hrsg. Bertelsmann Stiftung. Gütersloh. (Auch online unter https://doi.org/10.11586/2018006, Download 10.5.2018.)

Impuls Algorithmenethik #5: Dreyer, Stephan, und Wolfgang Schulz (2018). *Was bringt die Datenschutz-Grundverordnung für automatisierte Entscheidungssysteme?* Hrsg. Bertelsmann Stiftung. Gütersloh. (Auch online unter https://doi.org/10.11586/2018011, Download 10.5.2018.)

Adresse | Kontakt

Bertelsmann Stiftung
Carl-Bertelsmann-Straße 256
33311 Gütersloh
Telefon +49 5241 81-81216

Konrad Lischka
Taskforce Digitalisierung
Telefon +49 5241 81-81216
konrad.lischka@bertelsmann-stiftung.de

Ralph Müller-Eiselt
Taskforce Digitalisierung
Telefon +49 5241 81-81456
ralph.mueller-eiselt@bertelsmann-stiftung.de |

www.bertelsmann-stiftung.de

BertelsmannStiftung