TAB

BÜRO FÜR TECHNIKFOLGEN-ABSCHÄTZUNG
BEIM DEUTSCHEN BUNDESTAG

Alma Kolleck
Carsten Orwat

Mögliche Diskriminierung durch algorithmische Entscheidungssysteme und maschinelles Lernen – ein Überblick

Oktober 2020
Hintergrundpapier Nr. 24

Mögliche Diskriminierung durch algorithmische Entscheidungssysteme und maschinelles Lernen – ein Überblick

Das Büro für Technikfolgen-Abschätzung beim Deutschen Bundestag (TAB) berät das Parlament und seine Ausschüsse in Fragen des wissenschaftlich-technischen Wandels. Das TAB wird seit 1990 vom Institut für Technikfolgenabschätzung und Systemanalyse (ITAS) des Karlsruher Instituts für Technologie (KIT) betrieben. Hierbei kooperierte es seit September 2013 mit dem IZT – Institut für Zukunftsstudien und Technologiebewertung gGmbH sowie der VDI/VDE Innovation + Technik GmbH.

Alma Kolleck
Carsten Orwat

Mögliche Diskriminierung durch algorithmische Entscheidungssysteme und maschinelles Lernen – ein Überblick

TAB-Hintergrundpapier Nr. 24

Büro für Technikfolgen-Abschätzung
beim Deutschen Bundestag
Neue Schönhauser Straße 10
10178 Berlin

Telefon: +49 30 28491-0
E-Mail: buero@tab-beim-bundestag.de
Web: www.tab-beim-bundestag.de

2020

Umschlagbild (im Uhrzeigersinn): Mark Adams, Piotr Adamowicz,
Fabio Formaggio, photovibes; alle 123rf.com

ISSN-Print: 2199-7128
ISSN-Internet: 2199-7136

Inhalt

Zusammenfassung

Die Digitalisierung als zentrale Entwicklung der Gegenwart betrifft eine zunehmende Anzahl von Lebensbereichen – Arbeit, Bildung, Gesundheit, Handel, Kommunikation, Kultur, Verkehr, innere Sicherheit und viele weitere. Ein wesentliches Element digitaler Anwendungen stellen Algorithmen dar, die gleichsam ein konstitutiver Baustein dieser Entwicklung sind. Sie entscheiden über Kreditvergaben, unterstützen ärztliche Diagnosen und berechnen Fahrtwegeoptionen. Ein großer Teil algorithmischer Anwendungen bleibt für viele Nutzende weitgehend unbemerkt. Umso mehr Aufmerksamkeit erfahren Fälle, in denen algorithmische Entscheidungssysteme (AES) zum Nachteil von Einzelnen entschieden haben – insbesondere, wenn diese nachteiligen Entscheidungen gesellschaftliche Ungleichheiten spiegeln und fortschreiben. Dies kann etwa dann auftreten, wenn ein lernendes algorithmisches System potenzielle neue Mitarbeitende identifizieren soll und anhand der bisherigen Einstellungspraxis Männer gegenüber Frauen bevorzugt. Solcherlei Fälle werfen die Frage auf, ob vermeintlich objektive AES etwa das Risiko für soziale Diskriminierungen verändern. Anhand von Fallbeispielen fasst die Untersuchung den aktuellen Wissens- und Diskussionsstand zusammen.

Algorithmen – Definition und Funktionsweise

Algorithmen bilden einen Teil von Software, werden aber zugleich allgemein als Begriff für programmierte Verfahren verwendet, die aus einem bestimmten Input (meist Zahlenwerte) mittels einer genau definierten, seriellen Schrittfolge einen gewünschten Output berechnen. Sie analysieren Daten, ordnen Informationen nach Relevanz und gestalten Kommunikations- und Informationsprozesse. Eine häufige Unterscheidung gliedert Algorithmen danach, ob sie *regelbasiert* funktionieren oder aber *lernen*, also aus Trainingsdaten eigene Funktions- und Analyseregeln ableiten. Lernende Algorithmen bilden einen Teil *maschineller Lernsysteme* (ML), die auch mit dem Begriff der künstlichen Intelligenz (KI) beschrieben werden. Ein weiterer Bereich der künstlichen Intelligenz, der dem maschinellen Lernen häufig gegenübergestellt wird, sind sogenannte regelbasierte Systeme in Form von Expertensystemen. Algorithmen liefern dabei Ausgaben bzw. Schlussfolgerungen mithilfe von expliziten Regeln, überwiegend in Form von Wenn-dann-Regeln, und entsprechend aufbereiteten Wissensbeständen, die meist von Experten geschaffen worden sind. Sowohl regelbasierte als auch lernende Algorithmen können Teil von *algorithmischen Entscheidungssystemen* sein. AES vollziehen sowohl die Datenerfassung und -analyse als auch die Deutung und Interpretation der Ergebnisse und schließlich die Ableitung einer Entscheidung(sempfehlung) aus den Ergebnissen. AES dienen häufig der

Vorbereitung oder Unterstützung, teilweise auch als Ersatz menschlicher Entscheidungsprozesse.

In Bevölkerungsbefragungen zeigt sich, dass viele Menschen nur vage Ideen davon haben, was Algorithmen sind und wie sie funktionieren. Während die Befragten an Algorithmen und künstlicher Intelligenz insgesamt die Genauigkeit und Effizienz der Rechenvorgänge positiv hervorheben, sehen sie alleinige Entscheidungen von AES skeptisch. Diese Skepsis gegenüber AES ist aus der psychologischen Forschung unter dem Begriff der *Algorithm Aversion* bekannt. Insbesondere, wenn Menschen bei einem AES Fehler bemerkt haben, neigen sie dazu, dem AES weniger zu vertrauen als einem menschlichen Entscheider – selbst wenn dieser höhere Fehlerquoten aufweist als das AES. Zur Interaktion von Menschen und AES liegen bislang kaum Erkenntnisse vor. Einzelne Studien deuten darauf hin, dass Menschen, die beruflich mit AES interagieren, ihre Eingaben in das System teilweise gezielt modifizieren, um die Ergebnisse des AES an ihre eigenen Einschätzungen anzupassen.

Soziale und algorithmische Ungleichbehandlung

Ungleichbehandlungen sind in vielen Bereichen gesellschaftlich selbstverständlich, etwa bei Altersgrenzen für den Zugang zu bestimmten Gütern (Autofahren, Alkoholkonsum, Wahlrecht). *Ungerechtfertigte* Ungleichbehandlungen hingegen werden unter dem Begriff der Diskriminierung gefasst. Diskriminierung beschreibt folglich eine *Benachteiligung* von Personen, indem diese entweder ungerechtfertigt *ungleich* behandelt werden wie *gleiche* Personen oder ungerechtfertigt *gleich* behandelt wie *ungleiche* Personen. Soziale Benachteiligungen verlaufen häufig entlang von unterschiedlichen sozialen Positionen und betreffen dabei insbesondere solche Personen, die hinsichtlich ihres Alters, Geschlechts, ihrer Gesundheit, ihrer Kultur oder ihrer Hautfarbe von der jeweilig dominanten sozialen Gruppe abweichen.

Gesellschaftliche Ungleichbehandlungen können sich in komplexe AES einschreiben und dadurch potenziell eine Vielzahl von Personen betreffen. Algorithmische Ungleichbehandlungen geschehen häufig auf der Grundlage statistischer Zusammenhänge und anhand von Ersatzvariablen für Merkmale, die zwar von Interesse, aber statistisch nicht erfasst sind. Solche statistischen Ungleichbehandlungen finden auch außerhalb von AES Anwendung: Sie nutzen eine Ersatzvariable (wie etwa das Alter), um auf das Vorliegen anderer Merkmale (etwa Reife) zu schließen und anhand dieser den Zugang zu bestimmten Gütern zu begrenzen (z. B. Konsum alkoholhaltiger Getränke). Je nachdem, auf welche Merkmale sich die statistische Ungleichbehandlung richtet und welche Konsequenzen damit einhergehen, erscheint eine Ungleichbehandlung gerechtfertigt oder nicht. Wenn etwa Personen aus einer bestimmten Wohngegend lediglich aufgrund ihrer Adresse keinen Kredit erhalten oder arbeitssuchende

Frauen mit kleinen Kindern pauschal als schwer in den Arbeitsmarkt integrierbar gelten, erscheinen die zugrundeliegenden statistischen Verallgemeinerungen zumindest rechtfertigungsbedürftig und können im Konflikt mit gesellschaftlichen Grundwerten oder Gesetzen stehen.

Rechtliche Rahmung

Ungleichbehandlung durch komplexe AES stellt zwar eine vergleichsweise neue Herausforderung dar, doch es existieren bereits technologieneutrale rechtliche Grundlagen für die Vermeidung von Diskriminierungen, nämlich das Allgemeine Gleichbehandlungsgesetz (AGG), die Persönlichkeitsrechte nach Artikel 2 Absatz 1 in Verbindung mit Artikel 1 des Grundgesetzes (GG) und die Verordnung (EU) 2016/679[1] (Datenschutz-Grundverordnung). Das AGG definiert Voraussetzungen für das Vorliegen einer Diskriminierung, benennt geschützte Merkmale (also Eigenschaften, anhand derer keine Benachteiligung stattfinden darf) und schafft unter bestimmten Voraussetzungen die Möglichkeit der Beweiserleichterung für Betroffene. Die Persönlichkeitsrechte nach dem GG schützen die freie Entfaltung des Individuums sowie das Recht auf Selbstdarstellung. Statistische Verallgemeinerungen durch AES können diese Rechte tangieren. Die Datenschutz-Grundverordnung verbietet gänzlich automatisierte Entscheidungen mit rechtlicher Wirkung über Personen, legt Informationspflichten fest und benennt Voraussetzungen für die Pflicht zu einer Datenschutz-Folgeabschätzung.

Fallbeispiele für algorithmische Ungleichbehandlungen

Algorithmische Ungleichbehandlung sowie ihr Zustandekommen sind häufig für Betroffene schwer nachzuvollziehen. Vier Fallbeispiele illustrieren im vorliegenden Hintergrundpapier die Vielfalt der Lebensbereiche, in denen AES zum Einsatz kommen, die Herausforderung, die Ursache für die jeweilige Ungleichbehandlung zu detektieren, sowie die Schwierigkeit, zwischen gerechtfertigter und ungerechtfertigter Ungleichbehandlung zu unterscheiden – also zu bestimmen, wann eine Entscheidung sozial diskriminiert.

Das erste Fallbeispiel zur Ungleichbehandlung in der medizinischen Versorgung durch AES und ML macht deutlich, wie algorithmische Fehlschlüsse zustande kommen können, wenn wichtige Informationen nicht im AES hinterlegt sind. So berechnete ein in einem Krankenhaus eingesetztes AES für Patientinnen und Patienten mit Mehrfacherkrankungen und chronischen Krankheiten ein geringeres Sterberisiko (interpretiert als Indikator für den Bedarf für eine stationäre

1 Verordnung (EU) 2016/679 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten, zum freien Datenverkehr und zur Aufhebung der Richtlinie 95/46/EG (Datenschutz-Grundverordnung)

Behandlung) als für Patienten, die ausschließlich an einer Lungenentzündung erkrankt waren. Dieser Fehlschluss entstand, da das AES mit Daten trainiert worden war, in denen chronisch und mehrfach erkrankte Personen eine intensive medizinische Behandlung erhalten hatten und deshalb eine geringe Sterblichkeit aufwiesen. Die intensive medizinische Betreuung war in den Trainingsdaten jedoch nicht so hinterlegt, dass sie in die Berechnung des Sterberisikos einfließen konnte. Folglich ordnete das AES schwer erkrankten Patienten ein geringeres Sterberisiko und damit einen geringen Betreuungsbedarf zu.

Der österreichische Arbeitsmarktservice setzt seit 2018 ein AES ein, das Arbeitssuchende anhand ihrer soziodemografischen Daten hinsichtlich ihrer Arbeitsmarktnähe bewertet und somit einer von drei Gruppen (arbeitsmarktnah bis arbeitsmarktfern) zuordnet. Während im ersten Fallbeispiel die den Ausgaben des AES zugrundeliegenden Berechnungen erst rekonstruiert werden mussten, sind die Rechenvorschriften bei dieser Klassifizierung von Arbeitssuchenden öffentlich einsehbar. Wären sie nicht publiziert worden, hätte die Öffentlichkeit vermutlich weder erfahren noch kritisiert, dass Frauen mit Betreuungspflichten vom österreichischen Arbeitsmarktservice als eher arbeitsmarktfern klassifiziert werden. In der veröffentlichten Berechnungsvorschrift zeigt sich, dass das weibliche Geschlecht per se zu einem Abzug in der Arbeitsmarktnähe führt und Betreuungspflichten (lediglich bei Frauen) zu einem weiteren Abzug. Für eine Frau mit kleinen Kindern dürfte es folglich schwierig sein, durch das AES in die arbeitsmarktnahe Gruppe eingeordnet zu werden. Stellt dies eine Diskriminierung dar? Oder vielmehr eine Förderung, da der Arbeitsmarktservice die mittlere Gruppe am stärksten begünstigt? Diese Fragen lassen sich nicht durch eine Betrachtung des AES allein beantworten, sondern nur unter Einbeziehung des gesamten Entscheidungs- und Umsetzungsprozesses, also der sozialen Rahmenbedingungen.

Im dritten Fallbeispiel spielt ebenfalls die Zuordnung zu Gruppen eine zentrale Rolle und zwar in den Bereichen der Rechtsprechung und des Strafvollzugs. Ein mit dem Akronym COMPAS bezeichnetes Risikoabschätzungssystem bewertet Verurteilte hinsichtlich ihrer Rückfallwahrscheinlichkeit und kommt in den USA etwa dann zum Einsatz, wenn über die Aussetzung einer Bewährungsstrafe bzw. die Höhe einer Kaution entschieden wird. In der Hochrisikogruppe mit den Personen, denen das höchste Rückfallrisiko vorhergesagt wurde, zeigte sich im Rückblick, dass die Vorhersage bei vielen afroamerikanischen Straftäterinnen bzw. Straftätern falsch lag. So blieben 45 % von ihnen gesetzestreu, während unter den weißen Personen aus der Hochrisikogruppe lediglich 24 % gesetzestreu blieben. Dies deuteten einige Autoren als diskriminierend, während andere diese Interpretation nicht teilten und darauf verwiesen, dass die Bewertung stark von den jeweiligen Vorstellungen über Gerechtigkeit und Fairness abhängt.

Das vierte Fallbeispiel belegt ebenfalls Ungleichbehandlungen anhand der Hautfarbe. Vergleichende Studien verschiedener Gesichtserkennungssysteme zeigen, dass weibliche und dunkelhäutige Gesichter von den in den USA gängigen gewerblichen AES am häufigsten falsch erkannt werden. Solche Systeme kommen in der Auswertung visueller Daten im Bereich der Strafverfolgung, des Gesetzesvollzugs, der inneren Sicherheit und der Prävention zum Einsatz. Falsch-negative Ergebnisse (d. h., wenn ein Gesicht in einer Datenbank nicht wiedererkannt wird) schmälern dabei potenziell die Effizienz der eingesetzten Systeme, falsch-positive Ergebnisse (wenn ein Gesicht fälschlicherweise einem anderen aus der Datenbank zugeordnet wird) können für den Einzelnen unangenehme Folgen haben. Das Beispiel veranschaulicht, dass strukturelle Fehler von AES häufig zulasten von Bevölkerungsgruppen gehen, die sozial weniger begünstigt sind und die somit potenziell mehrfach benachteiligt werden.

Die Fallbeispiele führen vor Augen, dass Ungleichbehandlungen durch AES für die Betroffenen oft schwer nachvollziehbar sind. Häufig besteht auch keine Möglichkeit des Opt-outs (also der Verweigerung der Teilnahme). Zudem sind oftmals die sozialen Rahmenbedingungen des Einsatzes – somit also die Frage, wie und durch wen die Ergebnisse des AES umgesetzt werden – dafür entscheidend, ob die Ungleichbehandlung als diskriminierend wahrgenommen wird oder nicht und ob sie geltendem Recht widerspricht oder nicht.

Handlungsoptionen zur Diskriminierungsvermeidung

Eine Reihe von Vorschlägen zielt darauf ab, die Diskriminierungsrisiken von AES zu minimieren. Dabei stehen die Herstellung von Transparenz, eine Kontrolle und Evaluierung von AES sowie eine einheitliche Regulierung im Zentrum der Diskussion. So kann beispielsweise eine Kennzeichnungspflicht dazu beitragen, den Einsatz eines AES für die Betroffenen transparent zu machen. Eine risikoadaptierte Bewertung von AES kann gesellschaftliche Folgewirkungen ex ante abschätzen und je nach Kritikalität verschiedene Kontrollmaßnahmen etablieren, und eine Förderung des kollektiven Rechtsschutzes mit der Möglichkeit der Verbandsklage kann eine einheitliche Regulierung begünstigen. Diese Maßnahmen stellen nur einen Ausschnitt der derzeit diskutierten Ansätze dar, die darauf zielen, einen gesellschaftlichen und rechtlichen Umgang mit AES zu entwickeln, der Raum für Innovationen und Entwicklung bietet und zugleich Bürgerinnen und Bürgern Sicherheit vor Intransparenz und Diskriminierungen beim Zugang zu gesellschaftlich verfügbaren Gütern gewähren.

1 Einleitung

Dürfen Menschen mit einem unsteten Lebenswandel Ferienwohnungen anmieten wie andere Menschen auch? Oder ist es vertretbar, sie aufgrund ihrer Lebensweise von Vermittlungsplattformen für Urlaubsunterkünfte auszuschließen? Für Airbnb, Inc., Anbieter einer der weltweit bekanntesten Plattformen zur Vermittlung von Ferienunterkünften, scheint die Antwort auf diese Fragen klar: Menschen, die durch ihre Onlinespuren in einen Zusammenhang mit Sexarbeit, Neurotizismus, Bösartigkeit, Alkohol- oder Drogenkonsum oder dem Verfassen von Onlineinhalten mit negativer Sprache gebracht werden können, sollen die Vermittlungsplattform nicht nutzen können.[2] Um dies sicherzustellen, nutzt Airbnb seit 2017 das lernende System »Trooly«, das Onlineinhalte im Zusammenhang mit den Nutzenden der Plattform scannt und möglicherweise problematische Nutzende detektiert, sodass diese von der Nutzung ausgeschlossen werden können (Blunden 2020; Dickson 2020).

Das Beispiel enthält eine Reihe von Aspekten, die in diesem Hintergrundpapier beleuchtet werden sollen. So entscheidet ein *lernendes algorithmisches System* anhand von *Onlineinhalten*, die *systematisch gescannt* werden, über die anzunehmende *Reputation* einer Person und erstellt somit ein auf *statistischen Zusammenhängen beruhendes Profil* der Nutzenden. Die konkret von der Nutzung ausgeschlossenen Personen sind im Fall der Vermittlungsplattform in erster Linie Sexarbeiterinnen und -arbeiter, deren Tätigkeit in den meisten Gesellschaften sozial stigmatisiert ist (Vanwesenbeeck 2017; Weitzer 2018). Die Ungleichbehandlung betrifft folglich Personen, die bereits einen *niedrigen sozialen Status* aufweisen bzw. aufgrund ihrer Tätigkeit sozial geächtet sind. Aus Sicht des Plattformbetreibers ist es nachvollziehbar, dass dieser haftungsrechtliche Schwierigkeiten vermeiden möchte und somit vereinzelt Nutzer ausschließt. Gleichwohl hat Airbnb den von der Nutzung ausgeschlossenen Personen weder mitgeteilt, dass ein AES zum Einsatz kam, noch welche Informationen dieses verarbeitete und auch nicht, aufgrund welcher Inhalte es zum Ausschluss kam. Dieses Element der *Intransparenz* taucht ebenfalls an vielen Stellen in diesem Bericht auf.

Komplexe algorithmische Systeme sind Teil der Lebenswelt der meisten Menschen in Deutschland, sie kommen beispielsweise bei der Internetsuche, bei Produktempfehlungen, der Personalentwicklung, der Wettervorhersage, in automatisierten Haushaltsgeräten, automatischen Fahrassistenzsystemen und virtuellen Assistenten wie Alexa oder Siri zum Einsatz (Seeger 2017; Zweig/Krafft 2018, S. 208). Die breite Verfügbarkeit von algorithmischen Systemen einerseits

2 Die problematischen Eigenschaften einer Person sind im Original benannt als »neuroticism«, »badness«, »involvement with drugs and alcohol« und als »authoring online content with negative language« (Dickson 2020).

und ihre fehlende Anschaulichkeit andererseits sorgen dafür, dass sie zwar viele Vorgänge effizienter machen, ihr Wirken für die Nutzenden aber schwer nachvollziehbar erscheint: »Diese unsichtbaren Algorithmen, die die Schrauben und Rädchen des modernen Maschinenzeitalters bilden, haben zahllose Dinge ermöglicht, von Social-Media-Feeds bis zu Suchmaschinen, von der Satellitennavigation bis zu automatischen Musikempfehlungen. Sie sind ebenso Teil unserer modernen Infrastruktur wie Brücken, Gebäude und Fabriken. […] Sie sagen uns, was wir anschauen, was wir lesen und mit wem wir ausgehen sollen. Gleichzeitig üben sie eine geheime Macht aus: Sie verändern langsam und unmerklich, was es heißt, ein Mensch zu sein.« (Fry 2019, S. 14)

Die Ambivalenz in der Wahrnehmung komplexer algorithmischer Anwendungen und künstlicher Intelligenz im Allgemeinen bildet sich sowohl in der popkulturellen Beschäftigung mit diesem Phänomen (etwa im Kinofilm »Her« von Spike Jonze oder als selbstfahrendes Auto K.I.T.T. aus der US-Serie »Knight Rider«; Cave/Dihal 2019) als auch in der öffentlichen Meinung ab (dies zeigen repräsentative Bevölkerungsbefragungen wie etwa in Berg 2018; Fischer/Petersen 2018; Streim/Dehmel 2018). Im politischen Raum finden AES derzeit besonders große Beachtung. So thematisiert sowohl die Bundesregierung in ihrer KI-Strategie als auch die Datenethikkommission der Bundesregierung, die Enquete-Kommission Künstliche Intelligenz – Gesellschaftliche Verantwortung und wirtschaftliche, soziale und ökologische Potenziale des Deutschen Bundestages und die High Level Expert Group on Artificial Intelligence (Hochrangige Expertengruppe für künstliche Intelligenz) der Europäischen Kommission in ihren »Ethik-Richtlinien zum Einsatz von Künstlicher Intelligenz« und dem »Weißbuch zur Künstlichen Intelligenz«[3] den aktuellen Stand und mögliche Entwicklungsperspektiven.[4]

Parallel dazu widmen sich zahlreiche Forschungsprojekte den soziotechnischen Aspekten von (lernenden) Algorithmen, beispielsweise die geförderten Projekte »Assessing Big Data« (ABIDA, 2015–2019)und »Governance von und durch Algorithmen« (GOAL, 2019–2021) des Bundesministeriums für Bildung und Forschung (BMBF), das Projekt »Ethik der Algorithmen« der Bertelsmann Stiftung (2017–2019), das Forschungskonsortium AI Ethics Impact Group (ab 2019) ebenso wie das Projekt »Entscheiden über, durch und zusammen mit algorithmischen Entscheidungssystemen« des Hans-Bredow-Instituts (gefördert durch die Volkswagen Stiftung, 2019–2022). Die Untersuchung »Diskriminierungsrisiken durch Verwendung von Algorithmen« durch Carsten Orwat

3 Da das »Weißbuch zur Künstlichen Intelligenz« (Europäische Kommission 2020) erst kurz vor Redaktionsschluss des vorliegenden Hintergrundpapiers veröffentlicht wurde und einen ersten Anstoß zu einem breiteren politischen Dialog zur gemeinsamen EU-weiten Regulierung geben soll, wird es im Folgenden nicht genauer dargestellt.

4 Genauere Informationen zu den Strategien bzw. Expertengruppen finden sich unter https://www.ki-strategie-deutschland.de/home.html; https://www.bundestag.de/ausschuesse/weitere_gremien/enquete_ki; https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence (19.11.2020).

(2020) vom Institut für Technikfolgenabschätzung und Systemanalyse (ITAS) des Karlsruher Institut für Technologie (KIT) wurde von der Antidiskriminierungsstelle des Bundes gefördert und stellt eine wichtige Quelle für das vorliegende Hintergrundpapier dar. In der Studie wird eine Vielzahl von Fällen (vermeintlicher) Diskriminierung durch Algorithmen beleuchtet, die aktuelle Rechtslage zusammengefasst, Konzepte und Hintergründe von Diskriminierung diskutiert und Handlungsbedarfe und -optionen für Politik und Gesellschaft benannt. Einige Textstellen sind in etwas gekürzter Form direkt aus der Studie von Orwat (2020) übernommen und entsprechend gekennzeichnet.

Auch über die bereits gestarteten Projekte hinaus sind zukünftig weitere Forschungsergebnisse im Themenfeld zu erwarten. So hat das BMBF im Frühjahr 2019 eine Förderrichtlinie für Forschungsprojekte zur »Erklärbarkeit und Transparenz des Maschinellen Lernens und der Künstlichen Intelligenz« veröffentlicht, und auch das Bundesministerium für Wirtschaft und Energie fördert Projekte im Bereich der künstlichen Intelligenz. Die Bundesregierung (2019) stockte 2019 die Fördermittel für wissenschaftliche Forschung rund um die künstliche Intelligenz um 190 Mio. Euro auf.

Schließlich befasst sich auch eine Reihe von Intermediären zwischen Wissenschaft, Wirtschaft, Politik und Gesellschaft mit komplexen algorithmischen Systemen und maschinellem Lernen, beispielsweise die Plattform Lernende Systeme, das Deutsches Forschungszentrum für Künstliche Intelligenz, die Stiftung Neue Verantwortung, das Deutsche Institut für Normung (DIN) sowie zivilgesellschaftliche Organisationen wie AlgorithmWatch. Wissenschaftliche Fachgesellschaften, wie etwa die deutsche Gesellschaft für Informatik, spielen eine Rolle als Vermittler zwischen wissenschaftlichen und öffentlichen Diskursen.

Insgesamt befasst sich eine Vielzahl von Akteuren und Unterfangen mit komplexen AES und ihren gesellschaftlichen Möglichkeiten und Auswirkungen, sodass zu erwarten steht, dass sich die öffentliche und fachwissenschaftliche Diskussion um komplexe AES in Zukunft breit weiterentwickeln wird.

Beauftragung und Aufbau des Berichts

Vor dem Hintergrund dieser vielfältigen Entwicklungen im Bereich der künstlichen Intelligenz und der komplexen algorithmischen Systeme hat der Ausschuss für Bildung, Forschung und Technikfolgenabschätzung des Deutschen Bundestages das Büro für Technikfolgen-Abschätzung beim Deutschen Bundestag (TAB) beauftragt, eine zusammenfassende Auswertung der verfügbaren Studienergebnisse in Form einer Synopse zu erstellen. Ziel des Hintergrundpapiers ist es, einen konzentrierten Überblick über die aktuelle Debatte zu Diskriminierungsrisiken komplexer algorithmischer Systeme zu geben.

Dafür werden im Kapitel 2 grundlegende Definitionen, Eigenschaften und Aufgaben von AES und des maschinellen Lernens vorgestellt. Behandelt werden sowohl Perspektiven aus Forschung, Entwicklung und Anwendung als auch

Fragen bzw. Aspekte der öffentlichen Meinung zu bzw. Wahrnehmung von Algorithmen und künstlicher Intelligenz. Ein besonderes Augenmerk liegt auf der Frage, wie Menschen mit den Ausgaben komplexer algorithmischer Systeme umgehen und wie sie mit diesen interagieren.

In Kapitel 3 wird sich dem Bereich der Ungleichbehandlung und Diskriminierung sowohl einzelner Individuen als auch gesellschaftlicher Gruppen gewidmet. Dazu werden verschiedene Formen der Benachteiligung in technischen sowie nichttechnischen Kontexten beleuchtet und die rechtlichen Grundlagen für den Schutz vor einer möglichen Diskriminierung durch den Einsatz komplexer algorithmischer Systeme dargestellt.

Im Zentrum des vierten Kapitels stehen vier Fallbeispiele von Ungleichbehandlungen durch AES, die aus den Lebensbereichen Gesundheitsversorgung, staatliche Arbeitsvermittlung, Justizbereich, Personenerkennung im Sicherheits- und Verkehrswesen kommen. Für Deutschland existiert bislang kein öffentlich bekannter, gut dokumentierter Fall von algorithmischer Diskriminierung, sodass die Beispiele in erster Linie aus den USA sowie aus Österreich stammen (einige der eingesetzten Systeme, etwa im Bereich der Gesichts- und Personenerkennung, können potenziell auch in anderen Ländern zum Einsatz kommen).

Abschließend werden in Kapitel 5 in einem Ausblick Handlungsoptionen zusammengefasst, die in der wissenschaftlichen Debatte zum Umgang mit komplexen algorithmischen Systemen und ihren gesellschaftlichen Folgewirkungen diskutiert werden.

2 Algorithmische Entscheidungssysteme und maschinelles Lernen

2.1 Definitionen und Charakteristika von algorithmischen Systemen

Der Begriff Algorithmus geht auf den letzten Namensteil des persisch-arabischen Mathematikers Muhammad ibn Musa Al Chwarizmi (latinisiert Algorismi) zurück. Algorithmen bilden einen Teil von Software, werden darüber hinaus aber allgemein als Begriff für programmierte Verfahren verwendet, die aus einem spezifischen Input (wie Zahlenwerte) mittels einer genau definierten, seriellen Schrittfolge einen bestimmten Output berechnen (Vieth et al. 2017, S. 9). Dabei sind auch Verzweigungen möglich; idealerweise sind die Anweisungen so formuliert, dass keine Interpretationsspielräume offenbleiben. Sie können sowohl wenig komplexe Ergebnisse wie das Sortieren einer Zahlenreihe nach ihrer Größe als auch hochkomplexe Ergebnisse beispielsweise zur Wegeoptimierung bei der Nutzung verschiedener Verkehrsmittel berechnen. Algorithmen sind folglich klar formulierte, präzise beschreibbare, schrittweise Verfahren zur Lösung von Problemen (Saurwein 2019, S. 35). Wesentliche Gütemerkmale eines Algorithmus bestehen in der Korrektheit und Effizienz der Verarbeitung (Reichmann 2019, S. 135). Algorithmen in Programmen und Computersystemen analysieren Daten, ordnen Informationen nach Relevanz und gestalten Kommunikations- und Informationsprozesse. Bereiche, in denen Algorithmen weitgehend ohne menschliche Eingriffe entscheiden, sind der Hochfrequenzhandel an Börsen und automatisierte Internetwerbung (Saurwein 2019, S. 36).

Wenn ein Algorithmus in einer Programmiersprache formuliert wird, kann er als Programm oder Software automatisch auf einem Rechner ausgeführt werden. Dabei hängt seine Funktionsweise nicht ausschließlich von den zu verarbeitenden Daten ab, sondern auch vom Ausführungskontext, also von parallel verwendeten Hard- und Softwarekomponenten (Datenethikkommission 2019, S. 62). Um den Einsatz eines algorithmischen Systems zu bewerten, reicht es folglich nicht, die Berechnungsvorschriften und verwendeten Daten zu kennen, sondern auch die Rahmenbedingungen und Verwendungszusammenhang des Einsatzes. Einen weiteren relevanten Einflussfaktor für das Funktionieren eines Algorithmus bzw. der Software bilden die Parameter. Parameter ermöglichen es, eine Software an die jeweilig benötigten Einsatzkontexte anzupassen und zum Beispiel Darstellungen oder Filter für die jeweilig benötigte Anwendung zu definieren (anhand von Parametern können beispielsweise Schnittstellen zu anderen Systemen konfiguriert werden oder vom System definierte Arbeitsabläufe

in ihrer Reihenfolge verändert werden). Je mehr Parametereinstellungen einer Software veränderbar sind, desto flexibler ist diese einsetzbar und desto entscheidender sind die jeweiligen Parametereinstellungen (Datenethikkommission 2019, S. 62). Die Komplexität vieler Algorithmen, ihre Interaktion mit anderen Algorithmen bzw. Lernumwelten und ihre Einbindung in proprietäre Software gelten als zentrale Ursachen für die oft wahrgenommene Intransparenz von Algorithmen (Reichmann 2019, S. 144). Dabei gilt, dass zwar Computerprogramme, die Algorithmen umsetzen, unter Urheberrechtsschutz fallen, Algorithmen als reine Rechenregeln hingegen als nicht schutzfähig angesehen werden (Seeger 2017, S. 1).[5]

Es gibt verschiedene Ansätze zur Differenzierung von Algorithmentypen. Reichmann (2019, S. 140 f.) unterscheidet drei Kategorien von Algorithmen in Abhängigkeit davon, ob sie exogene Daten aufnehmen oder nicht (sensorische versus blinde Algorithmen) und ob sie lernen oder nicht (sensorisch-nichtlernende versus sensorisch-lernende Algorithmen). Anders als blinde Algorithmen, die einem klaren Skript folgen (z. B. in einer Waschmaschine ohne Sensoren z. B. für die Trommelfüllung), verwenden sensorische Algorithmen Daten, die sie von extern eingespeist bekommen (bei modernen Heizungsthermen etwa über die Raumtemperatur). Sensorisch-lernende Algorithmen arbeiten mit externen Daten und passen dabei ihr Ergebnis dem Feedback von außen an (etwa in der Bilderkennung).

Eine weitere, sehr häufig genutzte Differenzierung unterscheidet im Wesentlichen zwei Typen von Algorithmen, je nachdem, ob sie *regelbasiert* sind oder *selbstlernend*. *Regelbasierte Algorithmen* bestehen aus direkten und eindeutigen Anweisungen, die ein Mensch vorgegeben hat und die zumeist aus verschiedenen Schritten analog zu einem Flussdiagramm bestehen. *Selbstlernende Algorithmen*, also lernende Maschinen, erhalten vom Programmierenden Trainingsdaten und suchen darin nach Mustern und Zusammenhängen. Der/die Programmierende meldet dem Algorithmus zurück, welche Ergebnisse dem Auftrag entsprechen, sodass die Maschine die Schritte, die zur Erlangung eines optimalen Ergebnisses nötig sind, selbst finden und definieren kann bzw. soll.[6] Der Algorithmus leitet dabei aus der Beobachtung einer großen Anzahl von Fällen Regeln ab (Fry 2019, S. 24; Vieth et al. 2017, S. 10).

Beide Algorithmentypen bergen unterschiedliche Vor- und Nachteile. Während regelbasierte Algorithmen leichter nachvollziehbar sind, da sie auf menschlich formulierten Anleitungen (und damit menschlicher Logik) basieren, sind sie zugleich auf die Lösung solcher Probleme begrenzt, »bei denen Menschen wissen, wie sie die Anweisungen schreiben müssen« (Fry 2019, S. 24).

5 Eine detaillierte Darstellung der geltenden rechtlichen Regelungen zu Algorithmen und Software findet sich im noch unveröffentlichten TAB-Arbeitsbericht zu »Data-Mining – gesellschaftspolitische und rechtliche Herausforderungen«.

6 Im Fall des sogenannten überwachten maschinellen Lernens, das ein Teil des maschinellen Lernens ist.

Bei Problemen, die nicht mittels einer eindeutigen Anweisung zu lösen sind, zeigen selbstlernende Algorithmen große Fortschritte (etwa in der bildlichen Objekterkennung oder der Spracherkennung) – allerdings um den Preis, dass die selbstgelernten Lösungswege des Algorithmus für menschliche Beobachter kaum nachvollziehbar sind und somit oft »auch für die besten menschlichen Programmierer ein Mysterium« darstellen (Fry 2019, S. 24).

Algorithmische Entscheidungssysteme (AES, teilweise mit Bezug auf das englische Äquivalent Algorithmic Decision Making – ADM) basieren auf Algorithmen und erzeugen als Software auf der Basis von eingegebenen Daten (Inputvariablen) eine einzelne Ausgabe (Output) (Zweig et al. 2018, S. 185). Dabei umfasst eine algorithmische Entscheidungsfindung sowohl die Datenerfassung und -analyse als auch die Deutung und Interpretation der Ergebnisse und schließlich die Ableitung einer Entscheidung(sempfehlung) aus den Ergebnissen (Vieth et al. 2017, S. 10). AES dienen häufig der Vorbereitung oder Unterstützung, teils auch als Ersatz menschlicher Entscheidungsprozesse. Aspekte menschlicher Entscheidungsprozesse – wie etwa Ermessensspielräume – fallen aus dem algorithmischen Teil der Entscheidungsfindung heraus. Die verwendeten Algorithmen können sowohl sensorisch als auch blind, sowohl regelbasiert als auch selbstlernend sein.

Eine Form der AES stellen Systeme dar, die Menschen bewerten und dabei anhand einer Reihe von Charakteristika einer Person einen einzigen Wert zur Bewertung oder Klassifizierung ausgeben (auch als Scoring bezeichnet; Zweig/Krafft 2018, S. 209). Teilweise werden solche Scoringsysteme zur Vorhersage genutzt, etwa wenn eine Person anderen Personen mit ähnlichen Eigenschaften zugeordnet wird und anhand des Verhaltens der *ähnlichen Personen* auf zukünftiges Verhalten der *Zielperson* geschlossen wird. Solcherlei Art von Scoringverfahren bringen ähnlich wie andere Testerfahren neben richtig-positiven Ergebnissen (z. B. eine Person wurde richtigerweise in Gruppe A sortiert) und richtig-negativen Ergebnissen (z. B. eine Person wurde richtigerweise nicht in Gruppe A sortiert) auch falsche Ergebnisse hervor, entweder falsch-positive (z. B. Person A wurde fälschlicherweise in Gruppe A sortiert) oder falsch-negative (z. B. Person A wurde fälschlicherweise nicht in Gruppe A sortiert). Je nachdem, um welche Situation es sich handelt (also was Gruppe A jeweils ausmacht), sind die Folgen falscher Ergebnisse mehr oder minder tiefgreifend (beispielsweise je nachdem, ob Gruppe A all jene Individuen umfasst, die als Terroristen, als kreditwürdig, nicht weiter behandlungsbedürftig, rückfallgefährdet oder anderes bewertet wurden; Zweig/Krafft 2018, S. 212 f.). An diese fehlerhaften Ergebnisse und die Frage, wer von den fehlerhaften Ergebnissen besonders betroffen ist (z. B. Falsch-Positive oder Falsch-Negative bzw. die ganze Gesellschaft), knüpfen eine Reihe von Fairnesskriterien und Diskussionen um gerechte Ent-

scheidungen von AES an (für eine genauere Darstellung der Diskussion um verschiedene Gerechtigkeits- und Qualitätsmaße siehe Krafft/Zweig 2018 u. Zweig/Krafft 2018).

Lernende Algorithmen gehören zu einem Teilgebiet der Informatik, das als künstliche Intelligenz (KI) bezeichnet wird. Der Begriff ist nicht unumstritten, da das Verständnis, die Definition und die Messung von natürlicher Intelligenz als Grundlage der Ableitung des Begriffs künstliche Intelligenz keineswegs eindeutig oder einheitlich sind, sondern kontrovers diskutiert werden. In seiner vielzitierten Definition beschreibt Mainzer (2016, S. 3) künstliche Intelligenz folgendermaßen: »Ein System heißt intelligent, wenn es selbstständig und effizient Probleme lösen kann. Der Grad der Intelligenz hängt vom Grad der Selbstständigkeit, dem Grad der Komplexität des Problems und dem Grad der Effizienz des Problemlösungsverfahrens ab«. Beim maschinellen Lernen entwickeln Maschinen anhand von Beispieldaten Regeln und Modelle, die sie dann in einem zweiten Schritt auf neue Situationen anwenden. Beim *überwachten* Lernen erhält das zu trainierende System dann eine (menschliche) Rückmeldung, inwiefern die Ergebnisse der Aufgabenstellung entsprechen. Beim *unüberwachten* Lernen gibt es keine Rückmeldung, da es darum geht, »in einem großen, unstrukturierten Datensatz interessante und relevante Muster zu erkennen oder die Daten kompakter zu repräsentieren« (Beck et al. 2019, S. 10); es ist also in vielen Fällen gar nicht klar, wie das richtige Ergebnis aussieht, bzw. das Ergebnis lässt sich nur beschränkt von den Nutzenden auf seine Richtigkeit beurteilen.

2.2 Zielstellung und Entwicklung von algorithmischen Entscheidungssystemen

Algorithmische Systeme erfüllen in erster Linie vier Hauptaufgaben, nämlich Priorisieren, Klassifizieren, Kombinieren und Filtern. Algorithmen *priorisieren* Inhalte je nach Aufgabe nach angenommenen Interessen und Präferenzen (z. B. Suchergebnisse), nach festgelegten Parametern (z. B. Wegstrecke bei Navigationssystemen) oder nach anderen Gesichtspunkten (z. B. Ähnlichkeit bei der Gesichtserkennung). Sie *klassifizieren*, wenn sie unter einer abgeschlossenen Menge von Möglichkeiten die wahrscheinlichste ermitteln und danach etwa Nutzer als mögliche Interessenten einordnen, Spam-E-Mails erkennen und filtern oder handschriftliche Notizen als einzelne Buchstaben identifizieren und in Maschinenschrift überführen. Wenn Algorithmen *kombinieren*, fügen sie Verbindungen zwischen verschiedenen Informationen zusammen, etwa für Datingportale oder Produktempfehlungen auf Basis des Kaufverhaltens anderer Kundinnen und Kunden. Schließlich *filtern* Algorithmen aus verfügbaren Daten die (vermeintlich) relevanten Informationen heraus, etwa bei Spracherkennungssystemen, wie Alexa, Siri, Cortana und anderen, oder bei der visuellen Objekterkennung, beispielsweise in (teil)autonomen Fahrzeugen (Fry 2019,

S. 21 f.). Die meisten Algorithmen erfüllen mehrere der dargestellten Aufgaben gleichzeitig, etwa, wenn sie (im Rahmen der Navigation) eine Wegstrecke hinsichtlich ihrer Länge *priorisieren* und diese mit tagesaktuellen Informationen (wie etwa Staus oder Baustellen) *kombinieren* und dafür all jene Informationen *filtern*, die für die jeweilig geplante Strecke (Datum, Tageszeit) von Belang sind.

An der Entwicklung eines AES sind zumeist verschiedene Personen und Institutionen mit unterschiedlichen Expertisen und Aufgabenschwerpunkten beteiligt (Zweig et al. 2018, S. 190 f.). Oft verfügt der Auftraggeber über *Daten*, die zur Entwicklung des AES benutzt werden. Diese hat er entweder bereits gesammelt oder sammelt sie eigens zur Entwicklung eines AES und wählt ggf. im Folgenden besonders geeignete Daten aus. Diese Daten sowie Softwarepakete mit etablierten, viel genutzten Algorithmen dienen sogenannten Data Scientists (Datenwissenschaftlerinnen und -wissenschaftler) als Grundlage, um ein AES für den zuvor festgelegten Zweck zu entwickeln. Das entwickelte AES durchläuft mehrfache Tests und wird dann in den sozialen Zusammenhang eingeführt, für den es entwickelt wurde (z. B. durch Training der Anwenderinnen und Anwender). Möglicherweise findet nach der Inbetriebnahme des AES eine regelmäßige Rückmeldung an den Algorithmus bzw. an die Entwicklerinnen bzw. Entwickler statt, inwiefern die ausgegebenen Prognosen eingetroffen sind (sodass eine fortlaufende Anpassung des Algorithmus möglich wird). Folglich sind zumeist »in einem algorithmischen System verschiedene Akteure, beispielsweise in Form von Zulieferern, Betreibern oder Herstellern, für verschiedene Komponenten des Systems verantwortlich« (Datenethikkommission 2019, S. 62). Dadurch, dass die genutzten Trainingsdaten, die Softwarepakete, die Kombination der Softwarepakete und die Rückmeldungen an das AES jeweils unterschiedliche Urheberinnen und Urheber haben, entsteht eine Verkettung von Verantwortlichkeiten: »Diese große Gruppe von Beteiligten und die vielen Schritte im Prozess erschweren es, einzelne Entscheiderinnen und Entscheider zur Verantwortung zu ziehen und das beste AES zu gewährleisten.« (Zweig et al. 2018, S. 191)[7]

Häufige Probleme bei der Entwicklung eines AES umfassen fehlerhafte oder verzerrte Trainingsdaten, falsche Näherungsgrößen zur Messung einer nicht verfügbaren Zielgröße[8], die Auswahl von unpassenden Qualitätsindikatoren und eine fehlende Betrachtung, wie das AES in den jeweiligen sozialen Prozess eingebunden wird (Zweig et al. 2018, S. 191). Selbst wenn angestrebt wird, dass

7 Übersetzung TAB, im Original: »This large set of actors and the many steps of the process make it difficult to hold accountable individual decision makers and to guarantee the best ADM system.«

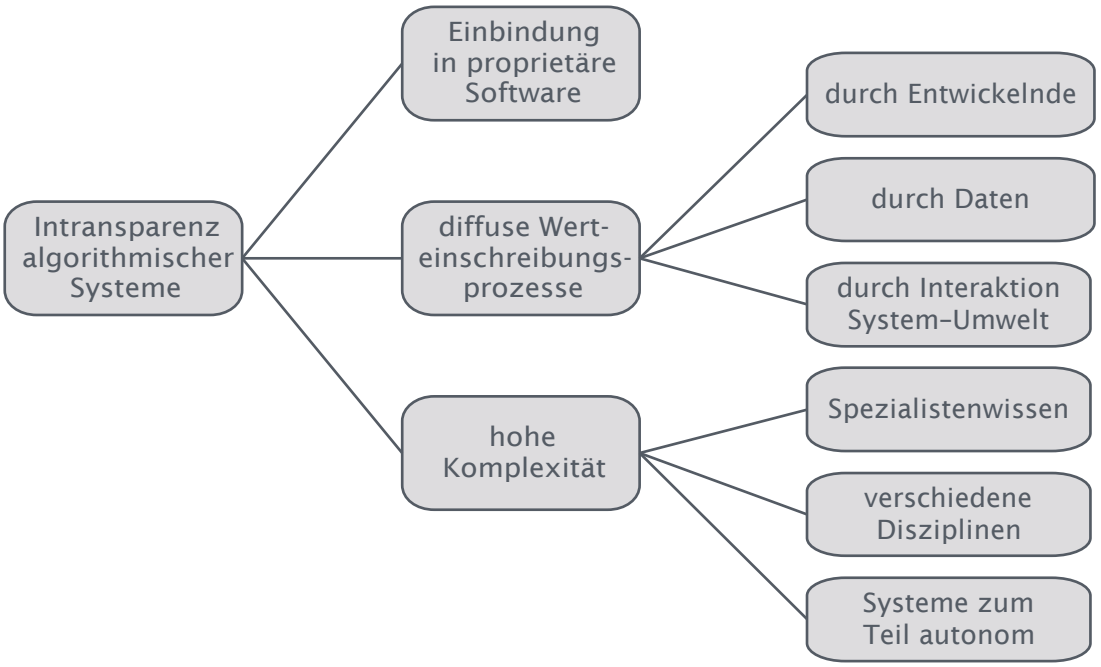
8 So lässt sich beispielsweise das Rückfallrisiko einer verurteilten Person nicht ex ante bestimmen. Um diese Lücke zu füllen, nutzen AES wie das in Fallbeispiel 3 in Kap. 4.3 diskutierte COMPAS-System eine Reihe von verfügbaren numerischen Variablen, um anhand dieser näherungsweise die Zielgröße das individuelle Rückfallrisiko zu berechnen.

ein AES parallel zu seinem Einsatz weiter trainiert wird, besteht bei einem fortlaufend evaluierten System – je nach Einsatz – die Herausforderung des asymmetrischen Feedbacks: Wenn die Ergebnisse eines AES beispielsweise dazu führen, dass Angehörige einer bestimmten Gruppe seltener auf Kaution freigelassen werden als die Mehrheitsbevölkerung, können Zugehörige zu dieser Gruppe auch seltener zeigen, dass das für sie errechnete Rückfallrisiko zu hoch lag und sie nicht rückfällig geworden wären (Rich/Gureckis 2019, S. 177; Zweig et al. 2018, S. 193).

Auch wenn viele einzelne Algorithmen, da sie größtenteils präzise formulierte Berechnungsschritte darstellen, grundsätzlich an sich in ihrem Vorgehen nachvollziehbar sind, wird häufig die fehlende Transparenz[9] algorithmischer Systeme moniert (Beck et al. 2019, S. 4; Reichmann 2019, S. 144; Saurwein 2019, S. 39; Zweig et al. 2018, S. 197 f.). So sind algorithmische Verfahren in Softwaresystemen für viele von ihnen Betroffene entweder gar nicht wahrnehmbar (weil ihre Verwendung für den Betroffenen nicht deutlich wird) oder ihr Funktionieren ist von außen nicht nachvollziehbar (etwa bei den Ergebnissen eines Kreditwürdigkeitsscoring). Die fehlende Nachvollziehbarkeit algorithmischer Systeme nicht nur für betroffene Laien, sondern oftmals auch für IT-Entwickelnde, speist sich aus einer Vielzahl von Quellen, die Abbildung 2.1 schematisch darstellt. Eine wesentliche Quelle von algorithmischer Intransparenz, die sich auch in einigen der in Kapitel 4 diskutierten Fallbeispiele zeigt, stellt die *Einbindung in proprietäre Software* und dem Geheimnisschutz sowie dem Schutz von Urheberrechten an Software dar, sodass die Rechenvorschriften, Trainingsdaten und Berechnungen zumeist nicht von Dritten (und Betroffenen) einsehbar sind (Reichmann 2019, S. 144; Zweig et al. 2018, S. 200). Insbesondere, wenn (vermeintliche) Diskriminierungen durch Algorithmen publik werden, zeigen sich zudem oft *diffuse Werteinschreibungsprozesse* in AES (Hagendorff 2019a, S. 54), die entweder durch die *Technikentwickelnden* selbst – bewusst oder unbewusst – in die Systeme übertragen werden oder aber durch verzerrte *Trainingsdaten* vom System erlernt oder in der *Interaktion* zwischen einem technischen System und der Umwelt entstehen (etwa wenn zur Berechnung einer Zielvariable Merkmale integriert werden, die bei bestimmten Gruppen systematisch anders sind als in der Gesamtbevölkerung; Beck et al. 2019, S. 15). Schließlich sind AES oft in hohem Maße *komplex* und dadurch auch für IT-Interessierte schwer zu durchblicken, sodass eine Nachvollziehbarkeit und Prüfung durch Dritte oft schwierig sind: »Künstliche Intelligenz gibt nicht vorrangig Algorithmen die Kontrolle, sondern denen, die sie entwickeln.« (Martini 2019, S. 49)

9 Transparenz wird im Folgenden verstanden als die Offenlegung des Codes eines AES gemeinsam mit der Dokumentation seiner Entwicklung, seiner Parameter sowie des Lerndatensatzes (sofern es sich um ein maschinell lernendes AES handelt) gegenüber Dritten, beispielsweise zur Prüfung oder Zertifizierung. Die Offenlegung muss nicht zwangsläufig gegenüber der Öffentlichkeit erfolgen (Castelluccia/Le Métayer 2019, II).

Abb. 2.1 Faktoren, die zur Intransparenz algorithmischer Systeme beitragen



Eigene Darstellung

Die Komplexität wird durch die Einbindung *verschiedener Spezialisten* und *Teilgebiete* in die Entwicklung algorithmischer Systeme weiter verstärkt.[10] Lernen Systeme *autonom* auf der Grundlage bereitgestellter Daten, lässt sich zudem oftmals schwer nachvollziehen, welche Lernprozesse auf welcher Grundlage zustande gekommen sind und welche Folgen sie in der praktischen Anwendung des Systems zeitigen (Hagendorff 2019a, S. 54; Saurwein 2019, S. 40).

Die Intransparenz komplexer algorithmischer Systeme begünstigt lückenhafte Verantwortungsstrukturen (Beck et al. 2019; Hagendorff 2019b; Jaume-Palasí/Spielkamp 2017; Saurwein 2019). Zudem befürchten viele Autoren und Autorinnen gesellschaftliche Folgewirkungen AES, die durch die unklaren Verantwortungsstrukturen begünstigt werden: »Wir geben zunehmend mehr Befugnisse an Algorithmen ab, und so werden sie rasch zu undurchsichtigen, aber entscheidenden Teilen der gesellschaftlichen Struktur.« (Sandvig et al. 2016, S. 4972[11]; ähnlich auch Hagendorff 2019a, S. 54) Befürchtet wird zudem, dass algorithmische Verfahren durch kontinuierliche Veränderungen in der soziotechnischen Interaktion grundlegende Wandelungsprozesse anstoßen könnten,

10 »Mit steigender Anzahl der Akteure sowie steigender Fragmentierung und Heterogenität verschlechtern sich die Voraussetzungen für kollektives Handeln, gemeinsame Entscheidungen und für kollektive Selbstregulierungsinitiativen, weil die freiwillige Einführung branchenübergreifender Mindeststandards erschwert wird« (Saurwein 2019, S. 41).

11 Übersetzung TAB; im Original: »We increasingly delegate authority to algorithms and they are fast becoming obscure but important elements of social structure«.

die jedoch unterhalb der Wahrnehmungsschwelle stattfinden, über lange Zeit unbemerkt bleiben und nicht öffentlich thematisiert bzw. ggf. korrigiert werden (Sutmöller 2019, S. 158).

2.3 Algorithmen und künstliche Intelligenz in der öffentlichen Wahrnehmung

Was wissen Bürgerinnen und Bürger über Algorithmen und maschinelles Lernen, und wie nehmen sie die Chancen und Risiken dieser Technologien wahr? Bei der Auswertung einer repräsentativen Befragung, die das Institut für Demoskopie Allensbach im Auftrag der Bertelsmann Stiftung 2018 mit mehr als 1.000 Teilnehmenden durchführte, zeigte sich, dass »in Bezug auf das Thema Algorithmen in Deutschland Unkenntnis, Unentschlossenheit und Unbehagen« (Fischer/Petersen 2018, S. 4) vorherrschen. Zwar gaben 72 % der Befragten an, den Begriff Algorithmus bereits gehört zu haben, aber die allermeisten Befragten (56 %) sagten, dass sie kaum etwas über Algorithmen wissen.[12] Lediglich 10 % der Befragten kannten sich nach eigenen Angaben mit der Funktionsweise von Algorithmen genauer aus (Fischer/Petersen 2018, S. 14). Den Befragten wurden verschiedene Anwendungsbereiche von Algorithmen genannt, von denen die bekanntesten die Personalisierung von Onlinewerbung (mit 55 % Bekanntheitsgrad) sowie die Selektion von möglichen Partnern und Partnerinnen in Internetpartnerschaftsbörsen (52 %) darstellten. Wenig bekannt hingegen war die Anwendung von Algorithmen bei der Diagnose von Krankheiten (28 %) sowie die Schätzung des Rückfallrisikos für Straftäter (18 %; Fischer/Petersen 2018, S. 15).

Insgesamt offenbarte sich eine eher skeptische Grundhaltung gegenüber Algorithmen. So sahen 36 % der Befragten bei algorithmischen Entscheidungen mehr Risiken als Chancen (46 % waren unentschieden und 18 % sahen mehr Chancen als Risiken; Fischer/Petersen 2018, S. 17).[13] Befragt nach den konkreten Charakteristika von Algorithmen assoziierte etwa die Hälfte der Befragten Genauigkeit (53 %), Fortschritt (50 %) und Effektivität (49 %), doch ein Drittel hielt sie zugleich für unheimlich (37 %) und unverständlich (37 %) und befürchtete einen Kontrollverlust (35 %; Fischer/Petersen 2018, S. 19). Im Einzelnen befürchteten die Befragten, dass Programmierende von Algorithmen zu viel Macht über die Nutzenden haben (57 %), Unternehmen zu viele Daten über

12 In einer Befragung im Auftrag des Max-Planck-Instituts für Bildungsforschung (Kozyreva et al. 2020, S. 9) im September 2019 kannten sogar nur 58 % der Befragten den Begriff »Computeralgorithmen«. Demgegenüber gaben 86 % an, »mehr oder weniger« zu wissen, was der Begriff »Künstliche Intelligenz« beschreibt.

13 In einer europaweiten repräsentativen Bevölkerungsbefragung, die im September 2018, also 8 Monate nach der Befragung von Fischer/Petersen (2018), stattfand und von dieser in Methodik sowie Formulierung der Fragen leicht abweicht, sehen 40 % der befragten Deutschen mehr Vorteile durch Algorithmen und 24 % mehr Probleme, 36 % sind unentschlossen (Grzymek/Puntschuh 2019, S. 24).

Menschen sammeln (68 %) und dass die Algorithmen leicht manipuliert werden können (55 %).

Bemerkenswerterweise waren diese Befürchtungen bei Personen, die bereits vor der Befragung eine Vorstellung vom Funktionieren von Algorithmen hatten, deutlich ausgeprägter: »Das Wissen über die Funktionsweise von Algorithmen geht also nicht nur mit einem größeren Verständnis für ihre Vorteile einher, sondern auch mit einem geschärften Bewusstsein ihrer Risiken.« (Fischer/Petersen 2018, S. 21) Insgesamt bevorzugten es acht von zehn Befragten, wenn *Menschen* statt *Computer* Entscheidungen über sie treffen. »Diese Abneigung ist in allen Gesellschaftsschichten vorhanden. In Deutschland herrscht gewissermaßen ein Konsens darüber, dass persönliche Entscheidungen über Menschen, mögen sie auch noch so fehlerhaft und subjektiv sein, automatisierten vorzuziehen sind.« (Fischer/Petersen 2018, S. 25)[14]

Befragt nach dem Begriff künstliche Intelligenz zeigten sich in einer weiteren repräsentativen Befragung durch Bitkom Research (Berg 2018) – trotz technischer Ähnlichkeiten und Überschneidungen mit dem Themenfeld Algorithmen – andere Assoziationen und Bewertungen. So geben 70 % der Befragten an, den Begriff künstliche Intelligenz zu kennen und seine Bedeutung erklären zu können – ähnlich dem Wert zur Kenntnis von Algorithmen. Mit 62 % sahen fast doppelt so viele Befragte künstliche Intelligenz eher als Chance gegenüber 35 %, die diese eher als Gefahr wahrnahmen (Berg 2018, S. 3 f.). In der zuvor Befragung von Allensbach zum Einsatz von Algorithmen hingegen war die Verteilung gleichsam spiegelverkehrt: Hier sahen 36 % mehr Gefahren und nur 18 % mehr Chancen (während 46 % unentschieden waren).

Hinsichtlich der praktischen Nutzung von künstlicher Intelligenz gaben 54 % der Befragten an, bereits digitale Sprachassistenten auf dem Handy genutzt zu haben, 37 % hatten automatische Übersetzungen angewendet. Die Befragten sahen breite Einsatzmöglichkeiten in der Unterstützung älterer Menschen (68 %), als Unterstützung für Ärztinnen und Ärzte in der Diagnostik (68 %) und in der Verwaltung (67 %). Gleichzeitig lehnten große Mehrheiten den Einsatz von künstlicher Intelligenz in der Betreuung von Kleinkindern (90 %) und im

14 Ähnliche Ergebnisse zeigte auch eine repräsentative Bevölkerungsumfrage über Big-Data-Praktiken, die in Szenarien zur Differenzierung von Krediten, Krankenversicherungstarifen, Beschäftigungsverhältnissen und Preisen im Handel eingesetzt werden würden. Bei der Umfrage wurden auch Einstellungen zu automatisierten Entscheidungen für die Differenzierungen abgefragt wurden. Beispielsweise lehnten 91,5 % automatisierte Entscheidungen, die allein durch den Computer und ohne menschliche Kontrolle erfolgen, ab. 79,7 % der Befragten verneinten, dass der Computer bessere Entscheidungen treffen könnte als ein Mensch, und 85,6 % stimmten zu, dass der Computer in diesen Fällen lediglich Empfehlungen abgeben und der Mensch immer entscheiden soll. 87,5 % stimmten zu, dass es verboten sein sollte, dass in solchen Fällen Computer die Entscheidungen alleine treffen. Insgesamt, fühlten sich 95,5 % unwohl, wenn Computer ohne menschliche Kontrolle solche Entscheidungen treffen (Orwat/Schankin 2018, S. 24 ff.). Ähnliche europaweit erhobene Ergebnisse finden sich in Grzymek/Puntschuh (2019, S. 25 ff.).

Beziehungsleben (etwa als Kommunikationspartner für einsame Menschen; 63 %) ab (Berg 2018, S. 5 ff.).

Die Gegenüberstellung der beiden repräsentativen Befragungen zeigt, wie stark sich die Ergebnisse je nach Formulierung der Sachverhalte und der Fragen sowie Antwortoptionen unterscheiden. So scheint der Begriff künstliche Intelligenz bekannter zu sein und positivere Assoziationen zu wecken als der Begriff Algorithmus. Dies ist umso erstaunlicher, als Algorithmen einen wesentlichen Baustein von Anwendungen darstellen, die unter dem Begriff der künstlichen Intelligenz zusammengefasst werden.

2.4 Mensch-Technik-Interaktionen: Wie gehen Menschen mit algorithmischen Entscheidungsvorschlägen um?

AES generieren aus einem zahlenbasierten Input über eine definierte Anzahl von Schritten, teilweise unter Hinzunahme weiterer definierter Parameter, einen konkreten Output, der nur in einigen Fällen direkt in eine Handlung überführt wird (etwa beim Hochfrequenzbörsenhandel). Üblicherweise sind AES in Entscheidungsstrukturen eingebunden, in denen Menschen die algorithmische Entscheidung erhalten, interpretieren und ggf. in eine Handlung umsetzen. Je nachdem, wie breit die menschlichen Entscheidungsspielräume im Umgang mit dem Output eines AES sind, unterscheidet die Datenethikkommission (2019, S. 17) zwischen *algorithmenbasierten*, *algorithmengetriebenen* und *algorithmendeterminierten* Entscheidungen. So sind *algorithmenbasierte* Entscheidungen all jene, die sich auf Informationen aus algorithmisch berechneten Verarbeitungsprozesses beziehen; *algorithmengetriebene* Entscheidungen entstehen auf der Basis algorithmischer Systeme und lassen dem menschlichen Entscheider nur noch wenig Spielräume; und *algorithmendeterminierte* Entscheidungen finden automatisiert statt, sodass in aller Regel keine menschliche Intervention in den Entscheidungsprozess vorgesehen ist.

Der Versuch, eine Trennlinie zwischen den verschiedenen Typen zu ziehen, dürfte in vielen Fällen zu strittigen Ergebnissen führen – denn wie breit Entscheidungsspielräume tatsächlich sind, ist von außen schwer zu ermessen. Oft wird angenommen, dass algorithmische Entscheidungen im Arbeitsalltag menschliche Ermessensspielräume sukzessiv schmälern: »Aus Mangel an Transparenz, Zeit oder Fachkenntnis ist es für die Nutzerinnen algorithmischer Entscheidungssysteme oft gar nicht möglich, Einzelfälle zu überprüfen. Zwischen einem teilautomatisierten und einem vollautomatisierten Prozess lässt sich in vielen Fällen nicht klar unterscheiden. Wenn ein Mensch die Entscheidung nicht mehr selbst mit Argumenten begründen kann, sondern sich auf den Output eines Computers verlässt, verschwimmt die Grenze zwischen menschlicher und algorithmischer Entscheidungsfindung.« (Vieth et al. 2017, S. 12)

Bislang fehlen eindeutige Ergebnisse dazu, *wie* Menschen mit den Ausgaben von AES umgehen und *welche Faktoren* ihren Umgang mit AES beeinflussen. Eine Ausnahme bildet die vielzitierte Studie von Dietvorst et al. (2015). In einer experimentellen Untersuchung ließen Dietvorst et al. (2015) über 2.000 Personen in verschiedenen Teilgruppen Vorhersagen auf der Basis algorithmischer oder menschlicher Expertise treffen. Die Studie verlief in zwei Phasen und in verschiedenen Teilgruppen. In der ersten Phase konnten die Probanden einiger Teilgruppen erstens nur die algorithmischen Vorhersagen sowie ihre Trefferquote, zweitens nur die menschlichen Vorhersagen und ihre Trefferquote oder drittens die menschliche und die algorithmischen Vorhersagen sowie ihre Trefferquote beobachten. Zudem gab es eine Kontrollgruppe, die weder menschliche noch algorithmische Vorhersagen beobachtete. Dabei stellten die Probanden fest, dass sowohl die algorithmischen Vorhersagesysteme als auch die menschlichen Probanden in ihren Prognosen teilweise falsch lagen. In der zweiten Phase sollten alle Probanden selbst Vorhersagen abgeben und konnten sich entscheiden, ob sie dabei die algorithmische oder die menschliche Prognose übernehmen wollten (eine richtige Prognose führte zur Erhöhung des Teilnahmeentgelts). Dabei zeigte sich, dass die Probanden den algorithmischen Prognosesystemen nicht länger vertrauten, sobald sie Fehler beim AES bemerkten.[15] Dies galt auch dann, wenn die menschlichen Vorhersagen deutlich schlechter abschnitten als die algorithmischen.

Es erscheint folglich, dass Menschen Algorithmen schon bei kleinen Fehlern nicht länger vertrauen, wohingegen menschliche Fehler (sogar wenn sie häufiger sind als beim Algorithmus) nicht zu einem Vertrauensverlust führen (Dietvorst et al. 2015, S. 125). Zugleich zeigen die Ergebnisse, dass Menschen eher bereit sind, Algorithmen zu nutzen, solange sie *nicht* merken, dass diese Fehler machen. Sind also die Algorithmen und ihre potenziellen Fehler für die Menschen nicht wahrnehmbar, kommt die von Dietvorst et al. (2015) beschriebene Algorithmenaversion nicht zum Tragen.

Gleichzeitig entwickeln Menschen Annahmen zu den Eigenschaften von Algorithmen, mit denen sie interagieren und stimmen ihr Handeln darauf ab (etwa, indem sie versuchen, den Algorithmus zu beeinflussen). Dies gilt insbesondere, wenn Menschen wissen, dass sie mit den Ergebnissen eines algorithmischen Auswahlsystems zu tun haben, dessen Funktionieren aber nicht nachvollziehen können, wie Bucher (2017) am Beispiel von Facebook zeigt. Der Algorithmus »EdgeRank« gestaltet den individuellen Newsfeed eines jeden Facebooknutzers anhand von dessen vermeintlichen Interessen. Damit entscheidet

15 Dieses Ergebnis galt für alle Probanden, die algorithmische Vorhersagen beobachtet hatten – entweder ausschließlich die algorithmischen oder sowohl die algorithmischen als auch die menschlichen Vorhersagen. Probanden, die lediglich menschliche Vorhersagen beobachtet hatten bzw. gar keine Beobachtungen zur Vorhersagequalität sammeln konnten (Kontrollgruppe), setzten in der zweiten Untersuchungsphase mehrheitlich auf die algorithmische Prognose (Dietvorst et al. 2015, S. 120).

der Algorithmus, welche Inhalte die Nutzenden sehen und worauf ihr Augenmerk beim Einloggen in das Netzwerk gelenkt wird. Einige der von Bucher (2017) befragten Nutzerinnen und Nutzern versuchten, sich so zu verhalten, dass der Algorithmus ihre Interessengebiete und Vorlieben leichter erkennen und berücksichtigen kann. Andere versuchten im Gegensatz dazu, möglichst undurchschaubar für den Algorithmus zu erscheinen und ihr Klickverhalten möglichst unvorhersehbar zu gestalten (Bucher 2017, S. 41).[16]

Diese Anpassung an algorithmische Analysesysteme gilt, so ist anzunehmen, auch über die Grenzen eines sozialen Digitalnetzwerks hinaus. Wird das Verhalten bei der Nutzung digitaler Inhalte durch die angenommene externe Analyse des eigenen Verhaltens beeinflusst, kann dies negative Wirkungen entfalten, wie Martini (2019, S. 50) ausführt: »Wer weiß, dass in der digitalen Reputationsökonomie sein Surfverhalten und seine Freunde die Chancen auf gesellschaftliche Teilhabe beeinflussen, wird im vorauseilenden Gehorsam auf potenziell score-schädigende Handlungen im Netz tendenziell verzichten.« Denke man diese Entwicklung weiter, bestehe für den Einzelnen die Möglichkeit der »subtilen Selbstzensur« und für die ganze Gesellschaft die Gefahr, dass »Risikoscheu, Konformitätsdruck und soziale Kontrolle wichtiger sind als Individualität, Innovationsfreude und Kreativität«.

Bucher (2017) zeigte, dass Facebooknutzerinnen und -nutzer, die wissen, dass sie mit einem Algorithmus interagieren, ihr Handeln daran ausrichten. Martini (2019) befürchtet, dass solche Anpassungen des eigenen Verhaltens an wenig transparente AES langfristig negative gesellschaftliche Folgen haben können.

Eine dritte Studie zur Interaktion zwischen Menschen und quantifizierten Entscheidungssystemen (Hannah-Moffat et al. 2009) zeigte erstens, dass eine Aneignung der Technologie durch die Nutzenden stattfindet (Angleichung der Systemurteile an die eigenen Urteile durch Modifikation der Eingaben in das System), und zweitens, dass die Charakteristika des Systems (vermeintliche Objektivität durch Quantifizierung) als Möglichkeit gesehen werden, die eigenen Urteile vor Kritik von außen zu schützen. Hannah-Moffat et al. (2009) untersuchten, wie Beschäftigte im kanadischen Strafvollzug mit den Ausgaben eines Risikomanagementsystems umgehen und zeigten, dass quantifizierte Entscheidungssysteme von den Beschäftigten geschätzt, zugleich aber als minderwertig gegenüber der eigenen Entscheidungskompetenz wahrgenommen wurden. Das in Kanada im Jugendstrafvollzug genutzte Risikomanagementsystem berechnet analog zu dem in Fallbeispiel 3 dargestellten System COMPAS das individuelle

16 Auch in der erwähnten Bevölkerungsumfrage zu Big-Data-Praktiken wurden mögliche Verhaltensanpassungen von Nutzerinnen und Nutzern untersucht, die sie angesichts von Differenzierungsentscheidungen zu Krediten, Krankenversicherungstarifen, Beschäftigungsverhältnissen oder Preisen im Handel unternehmen würden, wenn sie auf Basis von Auswertungen von Daten aus dem Internet erfolgen würden. Danach würden 69,7 % Maßnahmen zum Schutz der Privatsphäre treffen und 69,9 % würden zustimmen, dass sie darauf achten würden, nichts Nachteiliges über sich im Internet preiszugeben (Orwat/Schankin 2018, S. 26).

Risiko für zukünftige Gesetzesübertritte für die verurteilten Jugendlichen (etwa hinsichtlich der Rückfallwahrscheinlichkeit). Das System nutzt verschiedene Angaben zur jeweiligen Person (beispielsweise zur kriminellen Vorgeschichte, zum Elternhaus und zum sozialen Umfeld), um das jeweilige individuelle Risiko zu berechnen. Vor Einführung des Systems hatten die Beschäftigten die Risikostufe nach Begutachtung des/der Verurteilten sowie aufgrund der Aktenlage allein festgelegt.

In den Interviews wird deutlich, dass das Risikomanagementsystem von den Beschäftigten als Möglichkeit gesehen wird, ihr eigenes Urteilen gleichsam zu quantifizieren und objektivieren und damit für andere nachvollziehbar und unangreifbar zu machen. Eine der befragten Justizangestellten formulierte es folgendermaßen: »Ich denke, es gibt der Organisation und dem einzelnen Nutzer etwas Absicherung, dass ihre Planung und Entscheidungsfindung über das individuelle Risiko auf Fakten beruht.« (Hannah-Moffat et al. 2009, S. 397[17])

Die Befragten gaben zwar mehrheitlich an, dass das Risikomanagementsystem letztlich nur bestätige, was der *gesunde Menschenverstand* ebenfalls empfehle, die Quantifizierung und Externalisierung der Entscheidungsfindung böte jedoch eine Sicherheit, zum einen gegenüber anderen (wie Kolleginnen und Kollegen, Prozessbeteiligten und auch den Verurteilten) und zum anderen gegenüber sich selbst, wenn der jeweilige Fall kompliziert erscheine. Zugleich berichteten die Justizangestellten, dass die vermeintliche Objektivierung durch das Risikomanagementsystem vor allem die *Ausgabe* betreffe (also den Output als das zu präsentierende Ergebnis), die *Eingabe* jedoch stark von den Interpretationen des jeweiligen Bearbeitenden abhänge, etwa mit Blick auf die soziale Situation der Jugendlichen: »Informationen zu sammeln und zu bewerten und Empfehlungen für das Gericht oder die Strafvollzugsbehörden zu formulieren, beinhaltet eine Reihe von subjektiven Urteilen, die vom Vorwissen, den Erfahrungen, Werten und Überzeugungen des jeweiligen Betreuers beeinflusst sind, sodass verschiedene Betreuer bei derselben Faktenlage zu unterschiedlichen Ergebnissen gelangen können. Ermessensspielräume sind somit nicht abgeschafft, sondern hinter Risikokalkulationen verborgen und systematisiert.« (Hannah-Moffat et al. 2009, S. 401[18])

Ihre Ermessensspielräume nutzten die Justizvollzugsbeschäftigten nach eigenen Angaben insbesondere dazu, die Risikowerte des Systems nach unten

17 Übersetzung TAB; im Original: »I think it gives the organization and the individual user some comfort that their planning and their risk decision making is based on some evidence.«

18 Übersetzung TAB; im Original: »The exercise of gathering and assessing information and formulating recommendations for the courts or criminal justice institutions involves a range of subjective judgements that are informed by practitioners' personal knowledge, experience, values and beliefs, meaning that practitioners may even frame the same phenomena differently. Discretion is not eliminated, but becomes ›black boxed‹ and systematized, through risk calculations.«

oder oben zu korrigieren, indem sie die Eingaben in das Risikomanagementsystem anpassten (beispielsweise hinsichtlich der Faktoren, die Kriminalität statistisch begünstigen, wie Schulabsenz, Gewalterfahrungen in der Familie oder Drogenmissbrauch). Diese Anpassungen der Risikowerte durch gezielte Eingriffe nahmen die Nutzenden des Risikomanagementsystems in erster Linie bei sozial benachteiligten ethnischen Minderheiten (wie den First Nations) vor sowie bei Frauen, die kleinere Vergehen begangen hatten und nach Ermessen der Bearbeiterinnen und Bearbeiter vom System unverhältnismäßig hohe Risikowerte erhalten würden. Umgekehrt erhielten Straftäter mit schweren Sexual- und/oder Gewaltstraftaten von den Bearbeitern und Bearbeiterinnen häufig gezielt solche Bewertungen, die in einem hohen Risikowert mündeten. Damit korrigierten sie gleichsam einen von ihnen monierten blinden Fleck des Risikomanagementsystems, das Gewalt als möglichen Teil von Straftaten kaum berücksichtigt. Die Modifikation der Eingaben zur Anpassung der Risikowerte nahmen die interviewten Bearbeiterinnen und Bearbeiter vor, ohne dass sie dies gegenüber Vorgesetzten oder Kollegen dokumentierten, um somit bürokratische Vorgänge und Rechtfertigungszwänge zu vermeiden (Hannah-Moffat et al. 2009, S. 402 ff.).

Hannah-Moffat et al. (2009) zeigten, dass sich Nutzerinnen und Nutzer von Entscheidungssystemen mit dem System und seinen Ein- und Ausgaben kritisch auseinandersetzen und ihr Verhalten gezielt an das System anpassen, um Ergebnisse herbeizuführen, die ihrem Empfinden nach gerecht sind. Gleichzeitig schätzen sie die Verschiebung der Verantwortung auf eine vermeintlich objektive Entscheidungsinstanz und die Standardisierung und Quantifizierung der Entscheidungen. Einschränkend ist diesen Ergebnissen hinzuzufügen, dass sie auf Selbsteinschätzungen der befragten Justizangestellten beruhen. Zahlreiche Studien zeigen, dass Menschen ihr eigenes Handeln (im Rückblick) anders wahrnehmen und interpretieren als externe Beobachter (Pohl 2016). Vor diesem Hintergrund bleibt offen, ob die Einflussnahme der Beschäftigten auf das Risikomanagementsystem tatsächlich so einseitig erfolgt, wie von ihnen dargestellt, oder ob auch das Risikomanagementsystem das Handeln der Beschäftigten beeinflusst (ggf. ohne, dass ihnen dies bewusst wird – etwa, wenn sie im Prozess der Eingabe ihre eigene Risikoeinschätzung an das zu erwartende Ergebnis des Systems anpassen).

3 Ungleichbehandlung und Diskriminierung von Individuen und Gruppen

3.1 Definitionen und Charakteristika von sozialer Diskriminierung

Das deutsche Wort *diskriminieren* stammt vom lateinischen *discriminare* für unterscheiden ab. Die deutsche Bedeutung richtet sich gegenüber der neutralen lateinischen Wurzel jedoch auf mögliche negative Handlungsfolgen des Unterscheidens im Sinne einer *ungerechtfertigten Ungleichbehandlung*, die zwei Formen aufweisen kann: »Diskriminierung im negativen Sinn liegt bei einer ungerechtfertigten Ungleichbehandlung von Gleichem oder einer ungerechtfertigten Gleichbehandlung von Ungleichem vor.« (Beck et al. 2019, S. 7)

Doch wann ist eine Ungleichbehandlung gerecht(fertigt) und wann nicht? Wann ist dementsprechend eine (Un-)Gleichbehandlung diskriminierend? Die Antidiskriminierungsstelle des Bundes (ADS) beschreibt ungerechtfertigte (Un-)Gleichbehandlungen als *Benachteiligungen*. Diese liegen vor, »wenn eine Person eine weniger günstige Behandlung als eine Vergleichsperson erfährt, erfahren hat oder erfahren würde«.[19]

Sozialwissenschaftliche Perspektiven heben in ihrem Zugang zu Diskriminierung die Zugehörigkeit zu verschiedenen gesellschaftlichen Gruppen mit unterschiedlichen sozialen Stellungen hervor. So besteht etwa für Scherr (2016, S. 9) Diskriminierung »in der gesellschaftlichen Verwendung *kategorialer Unterscheidungen,* mit denen soziale Gruppen und Personenkategorien gekennzeichnet und die zur Begründung und *Rechtfertigung* gesellschaftlicher (ökonomischer, politischer, rechtlicher, kultureller) *Benachteiligungen* verwendet werden. Durch Diskriminierung werden auf der Grundlage jeweils wirkungsmächtiger *Normalitätsmodelle* und Ideologien Personengruppen unterschieden und *soziale Gruppen markiert,* denen der Status des gleichwertigen und gleichberechtigten Gesellschaftsmitglieds bestritten wird« (Hervorhebungen vom TAB).

Diskriminierung ist folglich eine soziale Praxis, die den Zugang zu bestimmten materiellen wie immateriellen Gütern anhand von (vermeintlichen) Gruppenzugehörigkeiten beschränkt. Dabei dient die Abweichung vom jeweils angenommenen Normalfall als Unterscheidungsmerkmal und damit Diskriminierungsanlass. Als typischer »Normalfall« gilt »der erwachsene, männliche, physisch und psychisch gesunde Staatsbürger, der zudem kulturell (Sprache, Religion, Herkunft) und im Hinblick auf äußerliche Merkmale (Hautfarbe) der

19 https://www.antidiskriminierungsstelle.de/DE/Beratung/FragenUndAntworten/faq_node.html (19.11.2020)

Bevölkerungsmehrheit bzw. der dominanten gesellschaftlichen Gruppe angehört« (Scherr 2016, S. 8). »Diskriminierung kann die Form einer Unterprivilegierung, einer Benachteiligung, eines sozialen Ausschlusses, einer schlechten Behandlung oder sogar einer physischen Vernichtung annehmen.« (Kessler/Mummendey 2007, S. 489) Eine häufige Grundlage für Diskriminierungen sind Vorurteile, also Antipathien oder eine negative, affektiv aufgeladene Haltung gegenüber einzelnen sozialen Gruppen oder Angehörigen dieser Gruppen.

Im Kontext von Informationstechnik sieht Hagendorff (2019a, S. 55) eine negative soziale Diskriminierung als gegeben an, »wenn die aus Datenverwertungsverfahren heraus entstehenden Differenzierungen als ungerecht angesehen werden und Handlungsentscheidungen an Persönlichkeitsmerkmalen orientiert werden, welche in keinem relevanten Zusammenhang mit jener Entscheidung stehen« (Hagendorff 2019a, S. 55). Dabei unterscheidet man üblicherweise weiter zwischen direkter und indirekter Diskriminierung. *Direkte* Diskriminierung ist dann gegeben, wenn Entscheidungen zuungunsten einer Person in Abhängigkeit von geschützten Merkmalen wie Geschlecht, sexuelle Orientierung oder Religionszugehörigkeit getroffen werden. *Indirekte* Diskriminierung hingegen liegt bei Entscheidungen vor, die nicht direkt an diesen Merkmalen anknüpfen, sondern an anderen Charakteristika, die eine Korrelation zu einem geschützten Merkmal aufweisen, z. B. Teilzeitbeschäftigung und weibliches Geschlecht oder nicht deutsch klingender Name und ethnische Herkunft (Hagendorff 2019a, S. 55).

3.2 Diskriminierung durch algorithmische Entscheidungssysteme

Wie gelangen verzerrende oder diskriminierende Einstellungen in AES? Dafür gibt es verschiedene Wege (Beck et al. 2019, S. 8 f.). Erstens ist es möglich, dass die Technikentwickelnden Voreingenommenheiten explizit oder implizit direkt in ein Verfahren oder technisches Artefakt einbauen. Zweitens kann es passieren, dass sich Wertannahmen nicht intendiert in Verfahren oder Artefakte einschreiben, etwa, wenn bei lernenden Verfahren Trainingsdaten verwendet werden, in denen gesellschaftliche Ungleichbehandlungen abgebildet sind. Das System lernt diese Ungleichbehandlungen und setzt sie (oftmals in gesteigerter Größenordnung) fort.[20] Möglich ist auch, dass die Wahl von Schwellenwerten

20 Ein Beispiel für eine nichtintendierte Ungleichbehandlung zeigte ein von Amazon entwickeltes AES. Dieses sollte eingehende Bewerbungen von möglichen neuen Mitarbeitern und Mitarbeiterinnen durchgehen und hinsichtlich ihrer Eignung bewerten. Das AES war mit den Bewerbungsunterlagen von Amazonmitarbeitern und -mitarbeiterinnen trainiert worden. Im praktischen Einsatz bewertete es männliche Bewerber deutlich besser als weibliche, was auf die Trainingsdaten zurückgeführt wurde, in denen Männer deutlich überrepräsentiert waren. In der Folge stoppte Amazon den Einsatz des AES (Dastin 2018).

oder Proxyvariablen (also Näherungsvariablen, wenn die Zielvariable nicht verfügbar oder messbar ist) zu nichtintendierter Diskriminierung führt (hierzu werden nachfolgend einige Beispiele skizziert; Barocas/Selbst 2016, S. 681; Hagendorff 2019a, S. 56 f.). Problematisch kann es auch sein, wenn in Trainingsdaten bestimmte Gruppen kaum oder unterrepräsentiert sind, etwa, weil zu ihnen aus Gründen der sozialen, technischen oder wirtschaftlichen Exklusion weniger Informationen vorliegen. Im Allgemeinen sind Daten, die zu geschäftlichen Zwecken gesammelt wurden, zumeist weniger repräsentativ und ausgewogen als etwa zu Studienzwecken erhobene Daten (Barocas/Selbst 2016, S. 684 ff.).

Grundsätzlich besteht eine Herausforderung algorithmischer Entscheidungsprozesse darin, dass sie auf (quantifizierbaren) Daten fußen, die für ein AES verarbeitbar sind. Dadurch ergibt sich zwangsläufig ein bestenfalls fragmentarisches Bild einer Person, das frei von sozialen Kontexten oder Reziprozität, also frei von einer gegenseitigen Bezugnahme von Urteilendem und Beurteiltem ist. Algorithmische personenbezogene Datensammlungen und -verarbeitungen bergen folglich die Gefahr, ein Bild von einer Person zu erzeugen, das ausschnitthaft und verzerrt ist »und möglicherweise im offensichtlichen Widerspruch steht zu dem Bild der Identität, welches die Person oder die Personengruppe selbst von sich zeigt oder zeigen möchte« (Hagendorff 2019b, S. 128). Diese Lückenhaftigkeit in der Beurteilung von Personen anhand von ausschnitthaft verfügbaren Informationen über sie ist selbstredend kein neues Phänomen, das erst durch den Einsatz von Algorithmen oder elektronischer Datenverarbeitung zu Tage getreten ist. »Dennoch kann konstatiert werden, dass insbesondere durch das Paradigma der Big Data eine Art digitaler Positivismus entstanden ist, in welchem Daten beziehungsweise Datenauswertungen so interpretiert werden, als würden sie die Realität abbilden und transparent machen.« (Hagendorff 2019b, S. 129)

Zudem können verzerrte Entscheidungen algorithmischer Systeme potenziell eine große Anzahl von Personen betreffen – deutlich mehr als verzerrte Entscheidungen, die durch eine einzelne Person getroffen werden. Benachteiligungen lassen sich somit potenziell systematisieren und vervielfachen: »Im Gegensatz zu vorurteilsbehafteten Entscheidungen einzelner Menschen besteht bei algorithmischen Systemen […] die Gefahr, dass der einem System inhärente Effekt über eine skalenmäßig große Anwendung des Systems eine Breitenwirkung entfaltet, die einzelne menschliche Bearbeiter nie erreichen könnten.« (Datenethikkommission 2019, S. 167)

Eine weitere Herausforderung in algorithmischen Ungleichbehandlungen liegt in einer möglichen Selbstverstärkung bei benachteiligenden Verfahren durch mangelnde Transparenz und fehlende Rückmeldungen: »Wenn diskriminierte Personen nicht gegen die Entscheidung vorgehen, erhalten weder das Programm noch der Softwareentwickler eine negative Rückkopplung zur feh-

lerhaften Entscheidung. Die Lernerfahrung bestärkt dann das verdeckt fehlerhafte Entscheidungsmuster. Im schlimmsten Fall mündet die Entwicklung darin, dass sich die diskriminierenden Kriterien zementieren.« (Martini 2019, S. 57) Eine andere Ursache für eine sich verstärkende Diskriminierung kann daraus resultieren, dass eine Personengruppe (z. B. Afroamerikanerinnen bzw. Afroamerikaner) durch eine Sicherheitstechnologie (z. B. Gesichtserkennungssoftware) besonders häufig falsch-positiv markiert werden (also z. B. als verdächtig), woraufhin in dieser Personengruppe überproportional viele Kontrollen erfolgen und die relative Wahrscheinlichkeit, als Person aus dieser Gruppe bei Vergehen ertappt zu werden, gegenüber dem Bevölkerungsdurchschnitt steigt – das (Vor-)Urteil wird zu einer sich selbsterfüllenden Prophezeiung (Rich/Gureckis 2019, S. 177). Kann es aus strukturellen Gründen keine Rückmeldungen über das Funktionieren des Systems geben (beispielsweise, weil gegenüber den Betroffenen der Einsatz des AES oder seine jeweiligen Parameter nicht offen gelegt werden), sind die Möglichkeiten zum Durchbrechen dieser sich selbsterfüllenden Prophezeiung stark limitiert. Weitere Beispiele folgen in Kapitel 3.3 und 4.

Es gibt verschiedene Ansätze, lernende Verfahren davor zu bewahren, bestehende gesellschaftliche Verzerrungen zu übernehmen, etwa indem Lerndatensätze so bereinigt werden, dass keine soziale Diskriminierung erlernt wird (siehe auch Orwat 2020, S. 99 ff.). Allerdings geht die Löschung von Attributen oftmals mit einer Verminderung der Analysequalität einher. In jedem Fall lassen sich diskriminierungsvermeidende Maßnahmen nicht global entwickeln, sondern müssen für die jeweilige Anwendung neu konzipiert werden (Hagendorff 2019a, S. 60 f.). Schließlich sind auch die zugrundeliegenden Konzepte von Fairness, Diskriminierung und (il)legitimer Ungleichbehandlung kulturell und zeitlich variabel und fallabhängig: »Universelle, über mathematische Verfahren verbriefte Richtlinien welche sich in Form technischer Handlungsschritte manifestieren, um Systeme diskriminierungsfrei zu machen, können nicht gefunden werden.« (Hagendorff 2019a, S. 63)

3.3 Statistische Ungleichbehandlung und statistische Diskriminierung

Eine Form der technikbasierten Diskriminierung, die sich im Kontext komplexer algorithmischer Systeme – insbesondere bei lernenden Systemen – häufig zeigt, ist die statistische Diskriminierung. Sie stützt sich auf Ersatzinformationen und stellt eine Sonderform der sozialen Diskriminierung dar: »Dass jemand ein Merkmal aufweist, das die Grundlage für die Ungleichbehandlung bildet, ist dann nur deshalb von Bedeutung, weil es darauf hinweist, dass die Person, die Ziel/Objekt/Gegenstand der Ungleichbehandlung ist, zugleich ein *anderes*

Merkmal besitzt.« (Schauer 2018, S. 43[21]) Beispiele für statistische Ungleichbehandlungen anhand von Ersatzinformationen spielen eine wichtige Rolle in unserer Alltagswelt und werden dann widerspruchslos angenommen, wenn die der Ungleichbehandlung zugrundeliegenden Ziele gesellschaftlich breit akzeptiert sind. Beispiele für solche statistischen Ungleichbehandlungen anhand des Alters umfassen etwa den Zugang zu bestimmten Gütern (motorisierte Teilnahme am Straßenverkehr, Kauf alkoholischer Getränke) erst ab einem definierten Lebensalter, da ein jüngeres Lebensalter mit mangelnder Fähigkeit zur Verantwortungsübernahme assoziiert wird. Ebenso gelten in bestimmten Berufen, wie etwa bei Pilotinnen und Piloten im gewerblichen Luftverkehr strikte Altersgrenzen (Pilotinnen und Piloten ab 60 Jahren dürfen nur im Beisein von jüngeren Kolleginnen und Kollegen fliegen, ab dem Alter von 65 dürfen Pilotinnen und Piloten gar nicht mehr im gewerblichem Luftverkehr fliegen). Ein gestiegenes Alter gilt hierbei als Ersatzinformation für eingeschränkte Sinneswahrnehmung und langsamere Reflexe (Schauer 2018, S. 44).

Die Altersdiskriminierung wird meist nicht als problematisch wahrgenommen, solange sie ein gesellschaftliches legitimes Ziel verfolgt und der statistische Zusammenhang zwischen der Ersatzinformation Alter und der Zielinformation Verantwortungsbewusstsein/Sinnesstärke evident erscheint, selbst wenn sie auf den Einzelfall nicht zutreffen mag. Von Bedeutung ist auch, dass das Alter eine temporäre, veränderliche Variable darstellt, die alle Menschen betrifft (anders als ethnische Herkunft, Religion, sexuelle Orientierung etc.). Problematisch wird statistische Diskriminierung hingegen insbesondere dann, wenn sie Merkmale betrifft, die als gesellschaftlich besonders schützenswert gelten und die mit einer gegenwärtigen oder vergangenen gesellschaftlich nachteiligen Stellung assoziiert werden (Schauer 2018, S. 52).

Die Näherungsgrößen (z. B. Alter) werden oftmals genutzt, da die Zielgrößen (z. B. Verantwortungsbewusstsein) entweder nicht direkt messbar oder nur mit großem Aufwand zu erheben wären. *Statistisch* und betriebswirtschaftlich betrachtet sind all solche Näherungsvariablen sinnvoll, die die Vorhersagegenauigkeit verbessern – *sozial sinnvoll* und *ethisch vertretbar* sind sie deshalb jedoch nicht zwangsläufig. Dies illustrieren zwei kurze Beispiele: So bekam ein solventer Lehrer aus Hannover keinen Mobilfunkvertrag bei einem Anbieter, trotz wohl überdurchschnittlicher ermittelter Bonität bei der beauftragten Auskunftei, da er in einem unterdurchschnittlichen Stadtviertel von Hannover, nämlich Linden-Nord, wohnte (Zgoll 2019). Die Näherungsgröße »Wohngegend« zur Bestimmung der Zielgröße »Zahlungskraft« scheint in diesem (Einzel-)Fall keine valide Prognose abgegeben zu haben – der Lehrer entschied sich

21 Übersetzung TAB; im Original: »possession of the trait that provides the basis for the discrimination is of significance to the discriminator only because of what it indicates about the likelihood that the person who is the subject of the discrimination possesses another trait.«

übrigens für einen anderen Anbieter. Ein ähnlicher Fall, bei dem einem Finnen von einer Bank mit Hauptsitz in Schweden nach einer Onlineprüfung ein Kredit verwehrt wurde, endete hingegen vor Gericht (Orwat 2020, S. 50 f.). Die Ablehnung des Kredits basierte auf den Angaben des Mannes zu seinem Geschlecht, Wohnort, Alter und zu seiner Muttersprache. Da er männlich und finnischsprachig (nicht schwedischsprachig)[22] sowie in einer finnischen Wohngegend beheimatet war, die vom AES als unbekannte Gegend klassifiziert wurde, wies er drei Merkmale auf, die für das AES auf statistischer Basis mit Rückzahlungsschwierigkeiten im Zusammenhang standen. Die Ablehnung erfolgte auf dieser Grundlage automatisiert, ohne eingehende Prüfung weiterer Gesichtspunkte. Der Kläger bekam Recht; die Entscheidung der Bank, ihn vom Kredit auszuschließen, wurde als Mehrfachdiskriminierung gewertet (YVTltk 2018).

Beide Fälle illustrieren, dass die (zugeschriebene) Zugehörigkeit zu einer Gruppe von Merkmalsträgerinnen und -trägern dazu führt, dass andere mit diesem Merkmal korrelierte Eigenschaften (wie etwa eine höhere Kreditausfallrate bei finnischsprechenden Männern bzw. Bewohnern von Linden-Nord) auf den Einzelnen übertragen werden, ohne dass der/dem Einzelnen die Möglichkeit gegeben wird, diese Annahme zu widerlegen. »Wenn aber eine Person nicht nach ihrem tatsächlichen Verhalten, ihren Fähigkeiten und Eigenschaften, sondern nur gemäß einer mehr oder minder groben Klassifikation beurteilt wird, ist das Ergebnis zwangsläufig kein gerechtes Urteil, sondern eine besondere Form des Vorurteils.« (Schaar 2017, S. 77) In den beiden Fällen zur Kreditgewährung bezog sich die (potenzielle) Diskriminierung auf direkt erfasste Merkmale (Wohnort bzw. Muttersprache und Geschlecht). Oftmals erfolgt die Ungleichbehandlung jedoch (häufig nichtintendiert) anhand einer verwendeten Näherungsgröße, von der zuvor nicht bekannt ist, dass sie nicht in allen Bevölkerungsgruppen gleich verteilt ist. Beim maschinellen Lernen wird oft eine Vielzahl von Variablen als Näherungsgrößen in den durch Lernverfahren gewonnenen Modellen verwendet, wodurch tendenziell das Risiko von direkten und indirekten Diskriminierungen steigt (Orwat 2020, S. 31 ff.).

Ein jüngstes Beispiel analysierten Obermeyer et al. (2019). Sie zeigten, dass ein in den USA breit genutztes AES zur Erkennung von besonders behandlungsbedürftigen Patientinnen und Patienten hautfarbenbezogene Verzerrungen aufweist. Von der Anwendung des Systems sind nach deren Angaben ca. 200 Mio. Personen pro Jahr betroffen. So werden schwarze Patientinnen und Patienten gemäß der algorithmischen Empfehlungen erst dann in die intensivere klinische Betreuung aufgenommen, wenn sie deutlich kränker sind als weiße Patientinnen und Patienten (also mehr chronische Krankheiten und

22 In Finnland stellt das Schwedische neben dem Finnischen (Suomi) die zweite Amtssprache dar. In Finnland lebt eine Minderheit von schwedischsprachigen Finnen, die etwa 5 % der Bevölkerung stellen und in Finnland als sozial höher gestellt wahrgenommen werden (Dutton et al. 2016, S. 45; Volanen et al. 2006, S. 515).

schlechtere Vitalparameter aufweisen) (Obermeyer et al. 2019, S. 449). Bei der Analyse des Algorithmus zeigte sich, woher diese Ungleichbehandlung rührte: Als Prädiktor (Vorhersagevariable) für den Gesundheitsstatus bzw. die Notwendigkeit einer umfassenden medizinischen Betreuung wurden durch den Hersteller des Systems die Kosten gewählt, die die zu behandelnde Person in der jüngeren Vergangenheit im Gesundheitssystem verursacht hatte. Je höher diese Kosten, so die Logik des Herstellers des eingesetzten Programms, desto höher die Wahrscheinlichkeit, dass auch zukünftig hohe Kosten anfallen, was als Indikator dafür interpretiert wird, dass die betreffende Person genauerer medizinischer Betreuung bedarf. Hieraus resultierte die Verzerrung hinsichtlich der Hautfarbe: Schwarze Patientinnen und Patienten verursachen statistisch im Gesundheitssystem weniger Kosten als weiße und werden daher durch die Messgröße »verursachte Kosten« in ihrer medizinischen Behandlung systematisch benachteiligt (Benjamin 2019). Die Gründe für die geringeren Gesundheitskosten sind nicht eindeutig zu benennen – »es mag am Umgang, am Vertrauen oder an Vorurteilen liegen, irgendetwas führt im Zusammenspiel von schwarzen Patientinnen und Patienten und Gesundheitssystem zu einer geringeren Inanspruchnahme« (Obermeyer et al. 2019, S. 450[23]) durch schwarze Patientinnen und Patienten[24] – die Folgen hingegen schon, sodass eine korrekte Vorhersage der Kosten zu einer Benachteiligung von Menschen mit dunkler Hautfarbe im Gesundheitssystem führt.

Werden solche Fälle von statistischen Ungleichbehandlungen und Diskriminierungen erkannt, lässt sich nicht auf einen fallübergreifend gültigen Weg zurückgreifen, mittels dessen sich die statistische Ungleichbehandlung eliminieren lässt. Vielmehr muss in Abhängigkeit von der jeweils diskriminierten Gruppe, dem verwendeten Verfahren, den Kontextfaktoren des Einsatzes und den Konsequenzen der Entscheidungen des AES eine Lösung erarbeitet werden. Eine Umgehung der Diskriminierung führt dabei fast zwangsläufig zu materiellen und/oder immateriellen Kosten (etwa weniger valide Vorhersagen des Systems), deren Verteilung auf verschiedene Weise gestaltet werden kann (Schauer 2018, S. 50).

Insgesamt kann zwar festgehalten werden, dass algorithmische Entscheidungsprozesse das Versprechen bergen, nicht nur effizienter, sondern auch objektiver zu entscheiden als Menschen (Wagner 2019): »So sind Algorithmen im Grundsatz auch weniger anfällig, Betroffene zu diskriminieren, als ein Sachbearbeiter aus Fleisch und Blut, der bei seiner Entscheidung Tagesstimmungen, einem sinkenden Blutzuckerspiegel, Müdigkeit, Launen und Vorprägungen – auch Vorurteilen – zu widerstehen sucht.« (Martini 2019, S. 47) Gleichzeitig

23 Übersetzung TAB; im Original: »Whether it is communication, trust, or bias, something about the interactions of Black patients with the health care system itself leads to reduced use of health care.«

24 Eine wesentliche Ursache kann der Zusammenhang zwischen dunkler Hautfarbe und geringerem sozioökonomischem Status darstellen (Obermeyer et al. 2019, S. 450).

agieren AES aber nicht objektiv in einem wertfreien Raum, insbesondere dann, wenn sie anhand von verzerrten Daten trainiert werden. Wohin dies führen kann, illustriert das Beispiel der Gesundheitsversorgung für Afroamerikanerinnen und Afroamerikaner sowie die Fallbeispiele in Kapitel 4.

Durch neue maschinelle Lernverfahren und ihren Umgang mit großen Datenmengen besteht darüber hinaus die Möglichkeit, dass sich *völlig neue* Ungleichbehandlungen durch AES auftun: »Zu beachten ist, dass das Versprechen von maschinellem Lernen (speziell bei unüberwachtem Lernen) auch darin liegt, neue Muster in Daten zu erkennen, die sich Menschen bislang nicht erschlossen haben […] durch völlig neue Korrelationen entstehen neue Unterscheidungen und Verknüpfungen, neue Gleich- und Ungleichbehandlungen – und wir werden intuitiv gar keine Vorstellung mehr davon haben, ob dieser Verdacht gerechtfertigt ist.« (Beck et al. 2019, S. 12)

Welche rechtlichen Regulierungen mit Blick auf (statistische und andere) Diskriminierungen durch komplexe AES greifen, wird nachfolgend dargestellt.

3.4 Rechtliche Aspekte des Umgangs mit Diskriminierung von Individuen und Gruppen

Bezüglich des Umgangs mit Diskriminierungsrisiken durch AES sind im deutschen Recht drei Regelungen einschlägig: das AGG, die Persönlichkeitsrechte nach GG und die Datenschutz-Grundverordnung.

Das AGG gibt grundlegende Definitionen vor, indem es zwischen unmittelbarer und mittelbarer Diskriminierung unterscheidet (hier und im Folgenden Orwat 2020, S. 26 f., 87 f., 106 ff. u. 114 ff.). Nach § 3 Abs. 1 des AGG liegt eine *unmittelbare Diskriminierung* vor, »wenn eine Person wegen eines in § 1 genannten Grundes eine weniger günstige Behandlung erfährt, als eine andere Person in einer vergleichbaren Situation erfährt, erfahren hat oder erfahren würde.« Dazu zählt § 1 AGG die Gründe bzw. die sogenannten geschützten Merkmale auf, die »Rasse«[25] oder ethnische Herkunft, Geschlecht, Religion oder Weltanschauung, Behinderung, Alter und sexuelle Identität umfassen (Tab. 3.1). Nach § 3 Abs. 2 AGG handelt es sich um eine *mittelbare Diskriminierung*, »wenn dem Anschein nach neutrale Vorschriften, Kriterien oder Verfahren Personen wegen eines in § 1 genannten Grundes gegenüber anderen Personen in besonderer Weise benachteiligen können, es sei denn, die betreffenden

25 Im Folgenden wird der Begriff »Rasse« in Anführungszeichen gesetzt oder durch den Begriff »ethnische Herkunft« ersetzt, wenn er in den zitierten Originaltexten, wie z. B. in aktuellen Gesetztestexten oder englischsprachiger wissenschaftlicher Literatur, verwendet wird. Damit wird der Empfehlung der UNESCO (1951) sowie von Biologinnen und Biologen gefolgt, die auf die fehlende wissenschaftliche Fundierung des Begriffs hinweisen, siehe auch z. B. Wersig (2017, S. 42).

Vorschriften, Kriterien oder Verfahren sind durch ein rechtmäßiges Ziel sachlich gerechtfertigt und die Mittel sind zur Erreichung dieses Ziels angemessen und erforderlich.«

Tab. 3.1 Geschützte Merkmale

	Artikel 3 GG	§ 1 und andere AGG**	Erwg. 71 Datenschutz-Grundverordnung	Artikel 9 Datenschutz-Grundverordnung
»Rasse« oder ethnische Herkunft	ja	ja	ja	ja
Abstammung, Heimat, Herkunft	ja			
Geschlecht	ja	ja		
Sprache	ja			
politische Meinung bzw. Anschauung und sonstige Anschauung	ja		ja	ja
Religion und Weltanschauung	ja	ja	ja	ja
Behinderung	ja	ja		
Alter		ja		
Gewerkschaftszugehörigkeit	ja*		ja	ja
genetische Merkmale bzw. Anlagen und Gesundheitszustand	ja		ja	ja
biometrische Merkmale				ja
Sexualleben, sexuelle Identität bzw. Orientierung		ja	ja	ja

* Nach Artikel 9 Absatz 3 GG;
** Darin besteht eine abgestufte Verwendung der Merkmale, z. B. gilt »politische Weltanschauung« nicht im Zivilrechtsteil (Wersig 2017).

Weitere Kataloge mit geschützten Merkmalen finden sich in der Charta der Grundrechte der Europä schen Union (Grundrechtecharta) und der Europäischen Konvention zum Schutz der Menschenrechte und Grundfreiheiten (Europäische Menschrechtskonvention – EMRK), die auch noch »Vermögen« und »Geburt« umfassen sowie die offene Klausel »sonstiger Status« in der EMRK.

Quelle: Orwat 2020, S. 25

Mit dem AGG sollten grundsätzlich Entscheidungen über Personen aufgrund bestimmter, weit verbreiteter Generalisierungen und Differenzierungen nach besonders generalisierungsanfälligen Merkmalen, die zu Diskriminierungen führen können, unterbunden werden (Britz 2008, S. 4). So ist grundsätzlich die unmittelbare Verwendung der geschützten Merkmale nach § 7 Abs. 1 und § 19 Abs. 1 AGG auch bei Differenzierungsentscheidungen mithilfe von Algorithmen bzw. Computersystemen unzulässig. Ebenso ist die bei algorithmischen Entscheidungen häufige mittelbare Diskriminierung, d. h., wenn ein scheinbar neutrales Merkmal verwendet wird, aber Personen im Hinblick auf die geschützten Merkmale benachteiligt werden, nach § 3 AGG verboten, es sei denn, die Verwendung des Merkmals ist durch ein rechtmäßiges Ziel gerechtfertigt und die Mittel zur Erreichung des Ziels sind angemessen und erforderlich (Verhältnismäßigkeitsprüfung) (nach Ernst 2017, S. 1032).

Das AGG sieht für mögliche Betroffene von Diskriminierung in § 22 AGG eine Beweiserleichterung vor, indem die der Diskriminierung beschuldigte Partei die Beweislast dafür trägt, dass kein Verstoß gegen die Bestimmungen zum Schutz vor Benachteiligung vorliegt. Die Beweiserleichterung ist nach Ebert (2019, S. 3018) allerdings an drei Voraussetzungen geknüpft: Die Person, die behauptet, diskriminiert worden zu sein, muss nachweisen, dass sie anders behandelt wurde als andere Personen, und sie muss nachweisen, dass sie sich im Hinblick auf eines der geschützten Merkmale nach § 1 AGG unterscheidet. Ferner müssen von der Person Indizien erbracht werden, die mit überwiegender Wahrscheinlichkeit darauf schließen lassen, dass das in § 1 AGG genannte Merkmal ursächlich für die Diskriminierung war. Diese Anforderungen sind bezüglich algorithmischer Diskriminierungsrisiken *problematisch*: Die aus Sicht der Betroffenen schlechte Nachvollziehbarkeit der Wirkungsweisen von Algorithmen stellt die Betroffenen vor die Schwierigkeit oder sogar Unmöglichkeit, eine Benachteiligung durch Algorithmen darzulegen (siehe auch ähnlich Hacker 2018, S. 1169).

Statistische Diskriminierung stellt eine Form der Benachteiligung dar, die auf dem Vorliegen von Merkmalen basiert, die mit bestimmten anderen Merkmalen in Zusammenhang gebracht werden (etwa ein bestimmtes Wohngebiet mit einer generell verminderten Kaufkraft oder das biologische Geschlecht einer Person und ihre Kreditausfallwahrscheinlichkeit; Kap. 3.3). Aus verfassungsrechtlicher Sicht kann im Hinblick auf die Gleichheitssätze nach Artikel 3 GG bei der statistischen Diskriminierung ein Generalisierungsunrecht auftreten, etwa wenn atypische Personen nach einem Ersatzmerkmal (z. B. Alter) von bestimmten beruflichen Tätigkeiten ausgeschlossen werden, diese eigentlich noch ausüben könnten (Britz 2008, S. 2 ff.). Ebenso kann von einer Unvereinbarkeit mit der Einzelfallgerechtigkeit gesprochen werden, denn: »In Fällen statistischer Diskriminierung wird ein Mensch wegen eines bestimmten (Stellver-

treter-)Merkmals anhand stereotyper Personenvorstellungen beurteilt und behandelt, ohne dass seine tatsächlichen Eigenschaften gewürdigt würden.« (Britz 2008, S. 12)

Es geht also um die Ungleichbehandlung durch Außerachtlassung der Besonderheiten des individuellen Falls. Da dies, wie für die statistische Diskriminierung kennzeichnend ist, zur kostengünstigen Überwindung von Informationsdefiziten geschieht, und Risiken von Ungerechtigkeiten damit quasi in Kauf genommen werden, kommt es zu Abwägungssituationen zwischen den eigentlich nicht vergleichbaren Werten Effizienz und Gerechtigkeit (Gandy 2010, S. 36 f.). Diese können kaum durch technische oder organisatorische Verbesserungen gelöst werden, sondern bedürfen der gesellschaftlichen Abwägung und Festlegungen in politischen und rechtssetzenden Prozessen. Der allgemein verbindliche Ausgleich sollte eigentlich durch das Recht, insbesondere durch das AGG geschehen. Aufgrund der relativ unkonkreten, generalklauselartigen Ausnahmeregelungen des AGG kann dies allerdings als nicht gelungen angesehen werden (Britz 2008, S. 72). Für die jeweilige Einzelsituation und vor allem dann, wenn neue Formen und Anwendungen des Typs der statistischen Diskriminierung auftreten, muss die Frage der Legitimität immer wieder neu gestellt werden. Hierzu wären algorithmische Verfahren in politischen Prozessen und durch den Gesetzgeber nach der Verhältnismäßigkeit zu prüfen. Die Verhältnismäßigkeitsprüfung bezieht dabei (Britz 2008, S. 151 ff.) den legitimen Zweck, die Eignung, Erforderlichkeit und Angemessenheit der Verfahren ein.

Neben den Gleichheitsrechten des GG sind durch das Phänomen der statistischen bzw. der algorithmen- und datenbasierten Diskriminierung auch die verfassungsrechtlich gewährten Persönlichkeitsrechte, insbesondere die durch Artikel 2 Absatz 1 geschützte freie Entfaltung der Persönlichkeit, betroffen. Das Problem resultiert daraus, dass sich Bewertende durch ein oder mehrere Merkmale ein bestimmtes Bild über die betroffenen Personen machen. Die Betroffenen werden mit fremdgefertigten Konstruktionen ihrer Identität, d. h. mit Fremdbildern konfrontiert (Britz 2008, S. 179 f.; Fröhlich/Spiecker 2019). »Statistische Diskriminierung nimmt der betroffenen Person die Möglichkeit, sich dem Gegenüber selbst darzustellen und damit zu beeinflussen, welches Bild man sich von ihr macht. Stattdessen wird nahezu automatisch von der Feststellung statistisch signifikanter Merkmale auf bestimmte Eigenschaften einer Person geschlossen. Statistische Diskriminierung stülpt den Betroffenen vorgefertigte Persönlichkeitsbilder über, denen sie weitgehend wehrlos ausgeliefert sind.« (Britz 2008, S. 124 f.) Kommt es zudem zu Fehlurteilen bei der statistischen Differenzierung, wird den Betroffenen unberechtigterweise eine bestimmte Eigenschaft zugeschrieben, »ohne dass sie sich dagegen im Prozess der Entstehung dieses Persönlichkeitsbildes durch eine eigene (Gegen-)Darstellung hätten zur Wehr setzen können« (Britz 2008, S. 180).

Dadurch wird den Betroffenen *ihr Recht auf Selbstdarstellung* genommen, das sich aus dem Recht auf freie Entfaltung der Persönlichkeit herleitet.[26] Nach Britz (2008, S. 179 ff.) ist die Selbstdarstellung das Mittel des Individuums, darauf Einfluss zu nehmen, welches Bild sich andere Menschen von ihm machen. Zum Schutz der Persönlichkeitsentfaltung sind Diskriminierungsverbote folglich nicht nur aus den grundrechtlichen Gleichheitsgrundsätzen, sondern vor allem aus dem Recht auf Selbstdarstellung und der zugrundeliegenden Garantie der freien Entfaltung der Persönlichkeit hergeleitet. Diskriminierungsverbote sind so auch als Schutz vor unzulässigen Fremdbildern und Fremdzuschreibungen zu verstehen (Britz 2008, S. 200 ff.).

Eine dritte wichtige Rechtsgrundlage zum Schutz vor algorithmischen Diskriminierungsrisiken stellen das Datenschutzrecht und darin konkret das sogenannte Verbot von automatisierten Entscheidungen dar. Der Zweck der Vorschrift findet sich bereits in der Datenschutz-Richtlinie (DSRL, RL 95/46/EG) sowie dem alten Bundesdatenschutzgesetz (BDSG a. F.) und dient dort dem Schutz der menschlichen Individualität als Element des Rechts auf freie Entfaltung der Persönlichkeit und autonome Gestaltung des eigenen Lebens (Ernst 2017, S. 1030; Hoeren/Niehoff 2018, S. 53; Martini 2018).

Nach Artikel 22 Absatz 1 Datenschutz-Grundverordnung hat eine betroffene Person das Recht, nicht einer ausschließlich auf einer automatisierten Verarbeitung beruhenden Entscheidung unterworfen zu werden, die ihr gegenüber rechtliche Wirkung entfaltet oder sie in ähnlicher Weise erheblich beeinträchtigt. Aus der Datenschutz-Grundverordnung geht nicht eindeutig hervor, welche Arten von automatisierten Entscheidungen tatsächlich erfasst sind. Das kann nur indirekt aus dem Wortlaut der Rechtsnorm abgeleitet werden:

Zum einen sind dies automatisierte Entscheidungen, bei denen eine ausschließlich auf eine automatisierte Datenverarbeitung beruhende Entscheidung vorliegt. Dies wird so interpretiert, dass dies der Fall ist, wenn keine inhaltliche Bewertung und darauf gestützte Entscheidung durch eine natürliche Person stattgefunden hat oder die involvierte natürliche Person keine Letztentscheidungskompetenz hat (Busch 2018, S. 31; Ernst 2017, S. 1029 ff.).

Zum anderen sind diejenigen Arten von automatisierten Entscheidungen erfasst, die »rechtliche Wirkung entfalten oder die Person in ähnlicher Weise erheblich beeinträchtigen« (Scholz 2019, Artikel 22 Rn. 31–37 Datenschutz-Grundverordnung). Bestimmend ist demnach die Art der Wirkungen. Eine *rechtliche Wirkung* ist dann anzunehmen, wenn die Rechtsposition der betroffenen Person sich verändert, wie z. B. bei Kündigung eines Vertrages; und eine *erhebliche Beeinträchtigung* ist immer dann gegeben, wenn die betroffene Person in

26 Zur Ausgestaltung des Rechts auf freie Entfaltung der Persönlichkeit und dem Recht auf Selbstdarstellung siehe auch Britz (2007). Darin zeigt die Autorin, dass das Recht auf freie Entfaltung der Persönlichkeit und das Recht auf Selbstdarstellung durch das Recht auf informationelle Selbstdarstellung und den Diskriminierungsverboten umgesetzt werden sollen.

ihrer wirtschaftlichen und persönlichen Entfaltung erheblich gestört wird, wie z. B. bei Versagen eines günstigen Zinssatzes (Busch 2018, S. 33).

Liegen automatisierte Entscheidungen vor, die nach den genannten Artikeln der Datenschutz-Grundverordnung erlaubt sind, dann sind weitere Regelungen einzuhalten (Hoeren/Niehoff 2018, S. 54 f.; Weichert 2018, S. 131): Artikel 14 Absatz 2 lit. g Datenschutz-Grundverordnung regelt die Informationspflichten so, dass beim Ausführen einer automatisierten Entscheidungsfindung »aussagekräftige Informationen über die involvierte Logik sowie die Tragweite und die angestrebten Auswirkungen einer derartigen Verarbeitung für die betroffene Person« durch die verantwortliche Stelle für die betroffene Person zur Verfügung gestellt werden müssen. Die Auskunftsrechte der betroffenen Person werden in Artikel 15 Absatz 1 lit. h Datenschutz-Grundverordnung geregelt und sehen, mit gleichem Wortlaut, Information über die involvierte Logik und Auswirkungen vor. Aus dem Wortlaut »Tragweite und angestrebte Auswirkungen« könnte auch eine Pflicht zur Information über mögliche Diskriminierungsrisiken erfolgen, was aber weiter rechtlich zu konkretisieren wäre (Orwat 2020, S. 120 f. mit weiteren Nachweisen).

Zudem wird bei automatisierten Entscheidungen eine Datenschutz-Folgenabschätzung nach Artikel 35 Absatz 3 lit. a Datenschutz-Grundverordnung erforderlich, wenn eine »systematische und umfassende Bewertung persönlicher Aspekte natürlicher Personen, die sich auf automatisierte Verarbeitung einschließlich Profiling gründet und die ihrerseits als Grundlage für Entscheidungen dient, die Rechtswirkung gegenüber natürlichen Personen entfalten oder diese in ähnlich erheblicher Weise beeinträchtigen«. Dadurch wird deutlich, dass der Gesetzgeber den automatisierten Entscheidungen ein hohes Risiko zuschreibt. Die verantwortliche Stelle muss mit der Datenschutz-Folgenabschätzung diese Risiken vorab bewerten, ebenso, ob die Verarbeitungsvorgänge notwendig und verhältnismäßig sind, genauso wie die zur Bewältigung der Risiken geplanten Abhilfemaßnahmen (Artikel 35 Absatz 7 lit. b bis d Datenschutz-Grundverordnung). Werden hohe Risiken festgestellt, muss eine Meldung an die Aufsichtsbehörde erfolgen (nach Artikel 36 Datenschutz-Grundverordnung). Die Aufsichtsbehörde kann in diesen Fällen die Verarbeitung untersagen (nach Artikel 58 Absatz 3 lit. f Datenschutz-Grundverordnung).

Gerberding/Wagner (2019) plädieren dafür, zukünftig insbesondere Scoringalgorithmen – die einen Teilbereich des Profilings darstellen – stärker regulatorisch in den Blick zu nehmen und für diese ein Qualitätssicherungsrecht zu verankern. Scoringalgorithmen kommen etwa zur Beurteilung der Kreditwürdigkeit (z. B. bei der Schufa) zum Einsatz, sind von zentraler lebensweltlicher Bedeutung (etwa für die Anmietung einer Wohnung oder zur Gewährung eines Kredites), und es ist für den einzelnen kaum möglich, diesen auszuweichen (Gerberding/Wagner 2019, S. 116). Das Datenschutzrecht bietet nach Ansicht

der Autoren keinen ausreichenden Schutz vor Scoringalgorithmen, da Qualitätsmängel bei Scoringalgorithmen vorkommen können, ohne dass personenbezogene Daten verarbeitet werden (Gerberding/Wagner 2019, S. 117). Die Last eines qualitativ unzureichenden Scorings betrifft schwerpunktmäßig diejenigen, über die eine Vorhersage getroffen wird, und »die Prognoseunsicherheit sieht man dem bloßen Scoringergebnis nicht notwendig an« (Gerberding/Wagner 2019, S. 118). Daher sollten Scoringalgorithmen eigenständige Bausteine eines »im Entstehen begriffenen Algorithmen-Regulierungsrechts« sein (Gerberding/Wagner 2019, S. 118).

Aus der Perspektive des Antidiskriminierungsrechts regt Martini (2019, S. 237) zudem eine mögliche zukünftige Erweiterung der durch das AGG geschützten Merkmale an – »nicht zuletzt, weil die Vielfalt der Gruppenzuordnungen und Merkmale, die Anknüpfung für Ungleichbehandlungen bieten, in der digitalen Welt wachsen wird« (Martini 2019, S. 238).[27]

27 Martini benennt keine neuen aus seiner Sicht zu schützenden Merkmale, sondern sieht es als Aufgabe der Rechtsgemeinschaft, zu bestimmen, »welche Merkmale die Rechtsgemeinschaft als so sensibel einstuft, dass die Privatautonomie des Einzelnen ihre Grenze finden soll« (Martini 2019, S. 239). Weitere Möglichkeiten zur diskriminierungsvermeidenden Regulierung von AES finden sich in Martini (2019, S. 230 ff.).

4 Fallbeispiele: Ungleichbehandlung durch AES in verschiedenen Lebensbereichen

4.1 Fallbeispiel 1: Ungleichbehandlung in der medizinischen Versorgung durch AES und ML

Mit dem Einsatz algorithmischer Systeme im Gesundheitssystem sind vielfältige Hoffnungen verknüpft, etwa eine verbesserte Nutzung und Verknüpfung vorhandener Informationen, eine Steigerung der Effektivität und Effizienz von Diagnostik und Therapie und eine Verbesserung der individuellen Behandlung durch individualisierte Medizin. Gleichwohl kann der Einsatz von künstlicher Intelligenz im Gesundheitssektor bestehende Ungleichheiten zwischen Bevölkerungsgruppen innerhalb eines Landes sowie zwischen ökonomisch stärkeren und schwächeren Nationen weiter verstärken (Nordling 2019). Um Ungleichbehandlungen vorzubeugen, erscheint es unabdingbar, dass algorithmische Berechnungen, die wichtige medizinische Entscheidungen unterstützen sollen, transparent und nachvollziehbar gestaltet werden, damit mögliche Fehler und Fehlschlüsse aufgedeckt werden können.

Caruana et al. (2015, S. 1722 ff.) berichten von Ungleichheiten bei der Anwendung eines Maschinenlernsystems, das zur Vorhersage der Sterblichkeit bei Patientinnen und Patienten mit Lungenentzündung eingesetzt wurde. Dabei zeigte sich, dass das System solche Patientinnen und Patienten, die neben der Lungenentzündung auch an Asthma, chronischen Lungenerkrankungen und chronischen Brustschmerzen litten, eine *geringere* Sterblichkeitsprognose stellte. Dieser Befund mag zunächst überraschen, ließ sich jedoch bei genauerer Betrachtung der zugrundeliegenden Berechnungen dadurch erklären, dass in der Vergangenheit Patientinnen und Patienten, die nicht nur an einer Lungenentzündung, sondern zugleich auch an einer chronischen Lungenkrankheit wie Asthma erkrankt waren, zumeist direkt intensivmedizinisch und stationär behandelt wurden. Aufgrund der intensiven medizinischen Betreuung wiesen die Mehrfacherkrankten in den Trainingsdaten des maschinellen Lernsystems eine geringere Sterblichkeit auf. Das System lernte daraus fälschlicherweise, dass die genannten chronischen Erkrankungen das Sterblichkeitsrisiko senken. Im beschriebenen Fall ließ sich der Fehler in der Berechnung nachvollziehen und korrigieren, denn das getestete Entscheidungsunterstützungssystem zeichnet sich dadurch aus, dass es die Grundlage der jeweiligen Behandlungsempfehlungen zugänglich macht. Weniger transparente Systeme würden die Aufdeckungen ähnlicher Fehler erschweren oder verunmöglichen (Cabitza et al. 2017, E2). Der Fall illustriert jedoch beispielhaft, dass statistische Zusammenhänge, die von Algorithmen aus Trainingsdaten gelernt werden, auf Entscheidungs- und Handlungsprozessen beruhen können, die unabhängig vom AES vonstattengehen

und die deshalb nur mit entsprechendem Fachwissen und nur bei einer kritischen Evaluation von AES nachvollziehbar (und damit korrigierbar) werden. Würden ein Arzt oder eine Ärztin auf der Grundlage des maschinellen Lernsystems entscheiden, ohne den Fehler im Algorithmus zu bemerken, würden die am stärksten gefährdeten Patientinnen und Patienten fälschlicherweise zur ambulanten Behandlung nach Hause entlassen (Caruana et al. 2015, S. 1721).

Die Vorhersage der Sterblichkeitswahrscheinlichkeit dient nicht nur der Entscheidung über eine stationäre oder ambulante Behandlung, sondern kann ebenso die Entscheidung über eine Aufnahme in die palliativmedizinische Pflege unterstützen. Eine palliativmedizinische Begleitung kann Patientinnen und Patienten, die nicht mehr geheilt werden können, am Lebensende Schmerzlinderung und ganzheitliche individuelle Unterstützung ermöglichen. Um terminal Erkrankte rechtzeitig in die Palliativversorgung aufnehmen zu können, müssen diese frühzeitig erkannt und informiert werden.

Verschiedene AES setzen an diesem Punkt an und prognostizieren den Todeszeitpunkt einzelner Patientinnen und Patienten anhand deren medizinischer Daten, um so geeigneten Patientinnen und Patienten palliative Pflege als Teil der Sterbebegleitung anbieten zu können (Avati et al. 2018; Briseño 2018). Dabei existieren sowohl universitär entwickelte AES zur Prognose des individuellen Todeszeitpunktes, die keine kommerziellen Interessen verfolgen (wie in Avati et al. 2018 vorgestellt), als auch kommerzielle Produkte, wie etwa das AES des US-amerikanischen Unternehmens Aspire Health, Inc., das in 26 US-amerikanischen Bundesstaaten angeboten wird (Briseño 2018) und das neben dem AES auch medizinische und pflegerische Dienste anbietet. Aspire Health betont auf der eigenen Webseite[28], dass sich durch den Einsatz des AES Kosten einsparen ließen. »Das Kalkül ist nun, dass man sich teure Untersuchungen [oder Behandlungen] sparen kann, wenn man zu wissen glaubt, dass es um den Patienten ohnehin bald geschehen sei. Für jeden Patienten wird ein medizinisches Ablaufdatum errechnet, das ihn als Risikopatienten oder hoffnungslosen Fall ausweist. Im Klartext heißt das: Ein Algorithmus bestimmt, wie jemand ärztlich versorgt wird.« (Lobe 2017, o. S.) Tatsächlich kann es für schwer Erkrankte ohne Heilungschance von Vorteil sein, wenn sie ihr Lebensende palliativmedizinisch ambulant betreut zuhause verbringen können, statt im Krankenhaus eventuell immer neue Behandlungen mit unklarem Ausgang zu durchleben. Unklar ist jedoch, welche Behandlungen von Aspire Health jeweils als unnötig betrachtet

28 Zur Kostenersparnis findet sich dort beispielsweise folgende Passage: »Die Pflege schwer Erkrankter ist einer der wenigen medizinischen Bereiche, in denen sich die Qualität der medizinischen Betreuung für die Patienten steigern und zugleich die Kosten reduzieren lassen. Derzeit entstehen etwa 25 % der gesamten Ausgaben im Bereich der Krankenversicherung im letzten Lebensjahr.« (Übersetzung TAB; im Original: »Serious illness care is one of the few areas in healthcare where the quality of care for patients can be significantly improved while simultaneously lowering costs. Today, roughly 25 % of all traditional Medicare costs are spent in the last year of life.« [http://aspirehealthcare.com/healthplans/; 19.11.2020])

werden, wie der Algorithmus darüber entscheidet und inwiefern die Berechnung des Algorithmus für die behandelnden Ärzte und Ärztinnen transparent ist. »Eine Validierung ihres Algorithmus bleibt Aspire Health bisher schuldig. Es gibt keine öffentlichen Studien darüber, wie hoch die Kosteneinsparungen tatsächlich sind, inwiefern eine Identifikation von Palliativpatienten durch den Algorithmus tatsächlich dazu beiträgt, den Zugang zur Versorgung zu verbessern, oder wie gut der Algorithmus auch die ›richtigen‹ Patienten identifiziert.« (Briseño 2018, o.S.) Da Aspire Health zudem die genaue Funktionsweise des Algorithmus nicht preisgibt, lässt sich dieses – an einer für das Individuum wie die Gesellschaft zentralen Stelle eingesetzte – AES auf mögliche Ungleichbehandlungen oder Fehler nicht testen.

Der Gesundheitsbereich führt zusammenfassend einige der Hoffnungen und Befürchtungen vor Augen, die mit dem Einsatz von algorithmischen Entscheidungs- und Lernsystemen verknüpft sind, etwa hinsichtlich der Nachvollziehbarkeit algorithmischer Entscheidungen, der Ökonomisierung von Krankenversorgung und damit von für den Einzelnen lebenswichtigen Entscheidungen sowie der Einbettung in soziale und technische Handlungskontexte (Nordling 2019, S. 105).

4.2 Fallbeispiel 2: Algorithmus zur Klassifizierung von Arbeitslosen in Österreich

Die hohe Dichte von weitgehend einheitlich erfassten Informationen und die große Anzahl von beteiligten Personen machen den Bereich der Arbeitsvermittlung und der Zuteilung von Lohnersatzleistungen und Qualifikationsangeboten zu einem Einsatzgebiet, in dem AES und maschinelles Lernen potenziell die Effizienz von Entscheidungsprozessen erhöhen können (Bundesregierung 2018; Schwär 2019).

Ein jüngeres Beispiel einer Ungleichbehandlung sozialer Gruppen im Bereich der Arbeitsvermittlung stammt aus Österreich. Um seine Ressourcen effizienter einsetzen zu können, setzt der österreichische Arbeitsmarktservice (AMS) seit Ende 2018 ein algorithmisches System ein, das seine Mitarbeiterinnen und Mitarbeiter bei der Entscheidung unterstützen soll, welche Arbeitssuchenden welche Fortbildungsangebote erhalten. Ziel ist es, die vorhandenen Ressourcen insbesondere auf solche Arbeitslose zu konzentrieren, bei denen angenommen wird, dass sie am meisten bewirken (also am stärksten die Chancen der Arbeitslosen am Arbeitsmarkt verbessern).[29] Dazu teilt der AMS mittels des

29 Johannes Kopf, der Vorstandvorsitzende des AMS, führte dies in einem Interview mit der Zeitung »Der Standard« wie folgt aus: »Wir setzen derzeit öfter geförderte Beschäftigungsprojekte bei ganz Schwachen ein und sind dann oft unglücklich, dass wir zu sehr hohen Kosten im Vergleich relativ wenige Arbeitsaufnahmen bei dieser Personengruppe haben.« (Szigetvari 2018)

Algorithmus die Arbeitssuchenden in drei Gruppen ein, die nach den jeweiligen (angenommenen) Chancen am Arbeitsmarkt klassifiziert werden. Dabei umfasst die oberste Gruppe all jene Kundinnen und Kunden des AMS, von denen angenommen wird, dass sie mit über 66%iger Wahrscheinlichkeit binnen eines halben Jahres eine Beschäftigung aufgenommen haben werden (dies ist die Gruppe mit der größten angenommenen Arbeitsmarktnähe). Menschen, von denen angenommen wird, dass sie mit unter 25%iger Wahrscheinlichkeit innerhalb der nächsten 2 Jahre eine Anstellung gefunden haben, werden der untersten, arbeitsmarktfernsten Gruppe zugeordnet. Alle Arbeitssuchenden, deren geschätzte Wahrscheinlichkeit zur Arbeitsaufnahme zwischen 25 % und 66 % liegt, werden der mittleren Kategorie zugeteilt (Holl et al. 2018, S. 6). Die meisten Ressourcen sollen für diese mittlere Gruppe aufgewendet werden, da angenommen wird, dass die Förderung in dieser Gruppe am meisten bewirkt (Fröhlich/Spiecker 2019). Die Beraterinnen und Berater des AMS sollen die Gruppenzuteilung als Hinweis für die Entscheidung darüber nutzen, ob der oder die Arbeitslose Weiterbildungs- oder andere Qualifikationsangebote erhält oder nicht. Die letztgültige Entscheidung über die Zuteilung von qualifizierenden Maßnahmen treffen, wie der AMS-Vorstand Johannes Kopf im Interview betont (Wimmer 2018a), die AMS-Mitarbeiterinnen und -mitarbeiter: »Die Technik trifft keine Entscheidung, sondern weist nur die Arbeitsmarktchancen aus.«

Die Arbeitsmarktchancen sind dabei gemäß Berechnungsgrundlage des Algorithmus für Frauen pauschal schlechter als für Männer. Weiteren Abzug erhalten Frauen (anders als Männer) für etwaige Betreuungspflichten (Wimmer 2018b). Bekannt wurde dies aus der Methodendokumentation der mit der Entwicklung des Algorithmus betrauten Firma Synthesis Forschung GmbH (Holl et al. 2018). Von Nachteil ist neben weiblichem Geschlecht (Faktor -0,14) und Betreuungspflichten (-0,15) ein Alter jenseits der 30 (-0,13 für 30- bis 49-Jährige; -0,7 für über 50-Jährige), körperliche oder psychische Beeinträchtigungen (-0,67) und die Herkunft aus einem Nicht-EU-Staat (-0,05). Weitere Abzüge sind für einen Wohnsitz in einem ungünstigen Arbeitsmarktbezirk[30] möglich (Holl et al. 2018, S. 11). Die Dokumentation beschreibt als Bezugsgruppe des Algorithmus »die Gruppe der jungen Männer mit höchstens Pflichtschulabschluss und österreichischer Staatsbürgerschaft. Sie haben keine Betreuungspflichten, sind nicht gesundheitlich beeinträchtigt und befinden sich in einem [vorteilhaften] Arbeitsmarktbezirk« (Holl et al. 2018, S. 11).

Das mediale Echo auf den Algorithmus und die Veröffentlichung der Rechengrundlagen fiel teilweise sehr kritisch aus: Der AMS-Algorithmus sei »sexistisch« (Gučanin 2018) und »ein Paradebeispiel für Diskriminierung« (Wimmer 2018b). Gegen diese Kritik wehrt sich der AMS, indem er hervorhebt, dass

30 Der AMS (2019a, 2019b) teilt Österreich in 97 Arbeitsmarktbezirke ein, in denen die Arbeitslosigkeit zwischen rund 6 und rund 12 % liegt. Der Arbeitsmarktbezirk und die jeweils zuständige Geschäftsstelle richten sich nach dem Wohnort der/des Erwerbslosen.

die Letztentscheidung beim Berater liege. Eine Diskriminierung sei nur gegeben, wenn der Algorithmus allein entscheide (Fanta 2018; Wimmer 2018a). Ist die Abbildung einer Diskriminierung von Frauen und gesundheitlich Beeinträchtigten am Arbeitsmarkt in der Berechnungsvorschrift eines Algorithmus bereits diskriminierend? Oder dient die Abbildung der Diskriminierung vielmehr dazu, dieser entgegenzuwirken? Der Leiter des AMS sieht den Algorithmus im Sinne der letztgenannte Frage als Werkzeug, die schlechtere Stellung von Frauen am Arbeitsmarkt durch ein Mehr an Förderung auszugleichen (Bachner 2018). Entgegengesetzt argumentieren Fröhlich/Spiecker (2019, S. 92), die eine negative Bewertung der Zugehörigkeit zum weiblichen Geschlecht ablehnen und grundsätzlich infrage stellen, »dass das Geschlecht ein solides Differenzierungsmerkmal ist«. Allhutter et al. (2020, S. 7) sehen die Gefahr, dass am Arbeitsmarkt tendenziell benachteiligte Gruppen wie Menschen mit Behinderungen und Frauen mit Betreuungspflichten durch die Klassifizierung des AMS im Sinne einer sich selbsterfüllenden Prophezeiung tatsächlich zu »hoffnungslosen Fällen« werden können.

Der sozialen Praxis und den Rahmenbedingungen kommt bei der Anwendung des Algorithmus eine Schlüsselstellung zu, darin sind sich AMS-Vertreter und ihre Kritikerinnen und Kritiker einig: »Es muss berücksichtigt werden, wie der Algorithmus im AMS verwendet wird und welche Anweisungen es zum Umgang mit der Klassifikation gibt. Wichtig ist etwa, unter welchen Bedingungen und mit welcher Begründung BeraterInnen die automatische Klassifikation verändern können und ob adäquate Schulungen des AMS-Personals durchgeführt werden« (ITA 2019). Dass sich die Beraterinnen und Berater des AMS in den alltäglichen Arbeitsentscheidungen von den Vorgaben des Algorithmus lösen und eine/-n Arbeitslose/-n eigenmächtig einer neuen Gruppe zuteilen, wird vielerorts bezweifelt (Allhutter et al. 2020, S. 12; Fanta 2018; Fröhlich/Spiecker 2019, S. 93; Wimmer 2018c), eine Evaluation des Einsatzes steht noch aus. Der Algorithmus befand sich zunächst ab 2018 in einer Testphase, ab 2020 soll er regulär und österreichweit zum Einsatz kommen (Bachner 2018).

In Deutschland werden derzeit nach Angaben der Bundesregierung (2018, S. 4 u.14) bei der Bundesagentur für Arbeit automatisierte AES eingesetzt, die nicht auf künstlicher Intelligenz bzw. maschinellem Lernen beruhen. Die Anwendungsbereiche umfassen die Berechnung der Arbeitsmarktchancen der Kundinnen und Kunden, eine psychologische Begutachtung zu beruflichen Eignung sowie ein Vermittlungssystem, das die Profile Arbeitssuchender mit Stellenausschreibungen und Fortbildungsangeboten abgleicht (genauere Informationen liegen bislang nicht vor: Bundesregierung 2018; Matzat et al. 2019, S. 27 f.; Schwär 2019).

4.3 Fallbeispiel 3: COMPAS, ein US-amerikanisches AES im Justizvollzug

Im Mai 2016 erschien auf der US-amerikanischen Nachrichtenseite ProPublica (Angwin et al. 2016) ein Artikel mit dem Titel »Machine Bias« (sinngemäß Automaten mit Vorurteilen). Die im Artikel beschriebene Diskriminierung durch AES ist vermutlich eines der bekanntesten Beispiele über maschinelle Entscheidungssysteme, die gesellschaftliche Ungleichbehandlungen widerspiegeln und damit potenziell verfestigen. Im Zentrum des Artikels und der daran anschließenden öffentlichen Diskussion steht das Bewertungssystem »Correctional Offender Management Profiling for Alternative Sanctions« (COMPAS; sinngemäß Steuerungssystem des Justizvollzugs zur Beurteilung Straffälliger bei der Entscheidung über Ersatzstrafen) zur Risikoabschätzung im Strafvollzug. Die Software COMPAS, so der Vorwurf von Angwin et al. (2016), beurteilte das individuelle Rückfallrisiko von dunkelhäutigen Straftäterinnen und -tätern systematisch als zu hoch und das der hellhäutigen Straftäterinnen und -täter als zu niedrig.

Die proprietäre Software COMPAS der Northpointe Inc. (heute Equivant) wird in einer Vielzahl US-amerikanischer Bundesstaaten eingesetzt und dient dazu, im gesamten Prozess der Rechtsprechung und Rechtsdurchsetzung Entscheidungen zur Aussetzung von (Bewährungs-)Strafen und zu Kautionshöhen zu unterstützen sowie die Urteilsfindung der Richterinnen bzw. Richter zu erleichtern (Angwin et al. 2016). Dabei stützt sich COMPAS auf Daten zum sozialen, ökonomischen und familiären Hintergrund einer Person (finanzielle Situation, Gewalterfahrung, Wohnsituation etc.) sowie auf Persönlichkeitsmerkmale (Anpassungsfähigkeit, Stresstoleranz etc.), die teils anhand der Akte des/der Beschuldigten und teils durch Befragung erhoben werden (Northpointe 2015). Die Hautfarbe des/der Beschuldigten wird nicht vermerkt. Ursprünglich zielten solche AES darauf, zu erkennen, welche spezifischen Unterstützungsleistungen eine Person bei ihrer Freilassung (auf Bewährung oder nach Ende des Prozesses bzw. der Strafe) benötigt, in der justizialen Praxis dienen sie jedoch zumeist dazu, das Rückfallrisiko einer Person auf einer Skala von 1 (niedrig) bis 10 (hoch) zu prognostizieren (Angwin et al. 2016).

Angwin et al. (2016) verglichen die COMPAS-Risikobewertungen von über 7.000 Menschen, die zwischen 2013 und 2014 in Broward County (Florida) festgenommen wurden, mit den tatsächlichen Rückfallzahlen. Als Rückfall werteten sie eine Festnahme innerhalb der folgenden 2 Jahre nach der COMPAS-Erfassung. Dabei zeigte COMPAS in der Zusammenschau aller dokumentierter Vergehen eine Trefferquote von 61 %, das heißt, dass von den Personen, denen COMPAS ein hohes Rückfallrisiko attestiert hatte, 61 % tatsächlich innerhalb

der folgenden 2 Jahre rückfällig wurden.[31] Angwin et al. (2016) nahmen dabei die Vorhersagegüte von COMPAS insbesondere in Abhängigkeit von der Hautfarbe in den Fokus und kamen dabei zu drei zentralen Ergebnissen. Erstens stellten sie fest, dass die Trefferquote bei der Vorhersage eines Rückfalls bei weißen und schwarzen Beschuldigten in etwa gleich hoch lag (59 % bei Weißen und 64 % bei Schwarzen). Zweitens bemerkten sie, dass Schwarzen häufig ein hohes Rückfallrisiko prognostiziert wurde, ohne dass diese tatsächlich später rückfällig wurden. Drittens hoben sie hervor, dass weißen Beschuldigten eine besonders niedrige Rückfallwahrscheinlichkeit vorhergesagt wurde, die sich jedoch häufig nicht bestätigte. Afroamerikanerinnen bzw. Afroamerikaner werden, so Angwin et al. (2016), folglich in ihrem Rückfallrisiko systematisch und massiv überschätzt, was – je nach Richterin bzw. Richter und konkreter Entscheidungssituation – beispielsweise in höheren Kautionssummen oder einer höheren Ablehnungsrate von Bewährungsstrafen resultieren kann.

Die Analyse von Angwin et al. (2016) blieb nicht lange unwidersprochen. So kritisieren Flores et al. (2016) ein fehlendes Verständnis der Autoren für den tatsächlichen Einsatz von COMPAS und bemängeln zudem methodische Fehler in der statistischen Analyse. In ähnlicher Weise betonen Corbett-Davies et al. (2016) anhand der von Angwin et al. (2016) genutzten Daten, dass sich die ungleiche Verteilung der Risikowerte u.a. darauf zurückführen lässt, dass afroamerikanische Verurteilte insgesamt mit einer höheren Wahrscheinlichkeit rückfällig werden (52 %) als weiße Verurteilte (39 %). Der Algorithmus gibt also gesellschaftliche Realitäten wieder, wenn er afroamerikanischen Verurteilten ein höheres Rückfallrisiko attestiert.[32] Offen ist die Frage, ob er damit die bestehende gesellschaftliche Realität eines insgesamt benachteiligten gesellschaftlichen Status afroamerikanischer Menschen (etwa hinsichtlich Bildung, Haushaltseinkommen, Gesundheitszustand; Berres 2017; Endres 2014; Nordling 2019) zusätzlich verfestigt.

Hamilton (2019) kommt in einer Studie, die auf den von ProPublica verwendeten Daten der COMPAS-Software fußt, zu dem Schluss, dass COMPAS Frauen hinsichtlich ihrer Rückfallwahrscheinlichkeit systematisch ein zu hohes Risiko berechnet. So sind in der Kategorie mit dem höchsten Rückfallrisiko 75 % der Männer, aber nur 65 % der Frauen rückfällig geworden, in der mittleren Risikokategorie 58 % der Männer und 46 % der Frauen und in der niedrigsten Risikokategorie 33 % der Männer und 24 % der Frauen (Hamilton 2019, S. 149). Dieses Missverhältnis führt Hamilton (2019, S. 147) darauf zurück, dass das ge-

31 Analysen auf der Basis anderer Daten zeigen etwas höhere Trefferquoten von 67,0 % für weiße und 63,8 % für afroamerikanische Angeklagte bei der Rückfallvorhersage durch COMPAS (Dressel/Farid 2018).

32 Diese Rückfallquoten errechnen Corbett-Davies et al. (2016) auf Basis des gleichen Datensatzes, den auch Angwin et al. (2016) verwenden. Der Datensatz umfasst über 5.000 Fälle, für die mögliche Rückfälle innerhalb der folgenden 2 Jahre vermerkt sind.

schützte Merkmal Geschlecht nicht in die Risikoberechnung von COMPAS aufgenommen wird. Da verurteilte Frauen statistisch gesehen seltener rückfällig werden als Männer, würden Frauen durch die fehlende Berücksichtigung des Merkmals Geschlecht benachteiligt. Ob es aus rechtlicher Sicht möglich wäre, das Geschlecht einer Person in ihre algorithmische Risikobewertung aufzunehmen, ist jedoch unklar (Hamilton 2019, S. 147 u. 154). In einigen Staaten wie Pennsylvania oder Wisconsin fließt das Geschlecht in die Risikoberechnung mit ein; eine gesamtstaatliche Regelung fehlt bislang.

Da die Ergebnisse der in den USA weitverbreiteten Risikobewertungssoftwareprogramme (neben COMPAS existieren noch weitere) zumeist in der Rechtsentscheidung nicht offengelegt werden, haben Angeklagte bzw. Verurteilte kaum Chancen, gegen den individuell errechneten Risikowert vorzugehen (Angwin et al. 2016). Desmarais und Singh (2013, S. 2 u. 51 f.) kommen zu dem Ergebniss, dass die große Mehrheit der in den USA verfügbaren und eingesetzten Risikoanalyseprogramme zur Vorhersage von Rückfallrisiken nicht extern evaluiert wurde. Der Begründer von COMPAS veröffentlichte 2009 eine Evaluationsstudie seiner eigenen Software, in der er zu dem Schluss gelangt, dass es nur geringfügige bzw. keine Unterschiede in der Genauigkeit der Vorhersagen nach Hautfarbe und Geschlecht gibt (Brennan et al. 2009).

Wodurch kommen die unterschiedlichen Risikobewertungen in Abhängigkeit von der Hautfarbe zustande? Northpointe/Equivant legt seine Berechnungen nicht offen, da diese unter das Geschäftsgeheimnis fallen (Angwin et al. 2016). Allerdings lässt sich aus dem Handbuch für Anwender der Software schließen, dass neben der kriminellen Vorgeschichte die individuelle Lebenssituation (z. B. Drogenabhängigkeit), Kontextfaktoren (z. B. keine gesicherten Wohnverhältnisse) und Persönlichkeitsmerkmale (z. B. Reizbarkeit) in die Berechnung des Risikowertes einfließen (Northpointe 2015). Die Hautfarbe ist keines der erfassten Merkmale, allerdings korrelieren viele der erfassten Eigenschaften mit der Hautfarbe (etwa Armut, Arbeitslosigkeit und soziale Ausgrenzung [Angwin et al. 2016; Dressel/Farid 2018]; dies schließt an das in Kapitel 3.3 dargestellte Phänomen der statistischen Ungleichbehandlung an).

Dressel/Farid (2018) verglichen die Trefferquoten von COMPAS in der Vorhersage der Rückfallwahrscheinlichkeit mit der Vorhersagegenauigkeit erstens einer linearen Regression mit sieben Variablen und zweitens einer Gruppe von Menschen mit wenig oder keinen juristischen Vorkenntnissen ebenfalls anhand von sieben Variablen. Die Regressionsrechnung wies eine Trefferquote von 66,6 % auf (gegenüber 65,4 % von COMPAS, das nach eigenen Angaben 137 Merkmale berücksichtigt), die Vorhersagen von Laien auf der Basis von sieben Merkmalsausprägungen der Verurteilten erreichten eine Trefferquote von 67,0 % (Dressel/Farid 2018). Mit Blick auf diese Trefferquoten halten Dressel und Farid (2018, o. S.)[33] fest: »Diese Ergebnisse wecken

33 Übersetzung TAB; im Original: »These results cast significant doubt on the entire effort of algorithmic recidivism prediction.«

ernste Zweifel am gesamten Unterfangen der algorithmischen Rückfallvorhersage.«.[34]

4.4 Fallbeispiel 4: algorithmische Personenerkennung anhand visueller Daten in den USA

Während ihres Studiums am Massachusetts Institute of Technology (MIT) machte Joy Buolamwini (2018) erstmals die Erfahrung, dass sie von Gesichtserkennungssoftware aufgrund ihrer dunklen Hautfarbe nicht erkannt wurde. Um studienbezogene IT-Programme, die eine Gesichtserkennung der Studierenden voraussetzten, dennoch nutzen zu können, verwendete sie infolgedessen eine schlichte weiße Maske. Mit dieser wurde ihr Gesicht erkannt und sie konnte die Software entsperren und nutzen.

Vor dem Hintergrund dieser Erfahrungen veröffentlichte sie jüngst zwei Studien zur algorithmischen Gesichtserkennung (Buolamwini/Gebru 2018; Raji/Buolamwini 2019), in denen sie zeigt, dass gewerbliche Softwaresysteme weibliche dunkelhäutige Gesichter am häufigsten falsch kategorisieren. Von drei getesteten Gesichtserkennungsystemen wiesen alle eine deutlich schlechtere Erkennungsrate des Geschlechts bei weiblichen Gesichtern (positiver Vorhersagewert, PVW: 79 bis 89 %) und bei dunkleren Hauttypen (PVW: 78 bis 87 %) auf. Weibliche Gesichter dunkleren Hauttyps wurden daher am seltensten richtig erkannt (PVW: 65 bis 79 %) (Buolamwini/Gebru 2018, S. 9). In einer Follow-up-Studie untersuchten Raji/Buolamwini (2019), inwiefern die drei bereits zuvor getesteten Softwarelösungen in Reaktion auf ihre Untersuchungen[35] zwischenzeitlich verbessert wurden. Tatsächlich wiesen sie eine deutliche Verringerung der Fehlerquoten bei dunkelhäutigen weiblichen Gesichtern auf (18 bis 30 % weniger Fehler für diese Teilgruppe gegenüber dem vorherigen Test). Zwei weitere erstmalig getestete Systeme zeigten allerdings, dass hinsichtlich der Geschlechtserkennung »bedeutende Leistungsunterschiede bestehen bleiben« (Raji/Buolamwini 2019, o.S.[36]). So offenbaren diese Systeme Fehlerquoten von 23 bis 31 % bei weiblichen Gesichtern mit dunklerer Hautfarbe.

Da die algorithmische Zuordnung zu Geschlechtern lediglich ein erster Schritt in einer Reihe maschineller Auswertungsentscheidungen ist und diese

34 Lin et al. (2020) konnten die Ergebnisse von Dressel/Farid (2018) in einem ähnlichen Versuchsaufbau reproduzieren. Veränderten sie jedoch die Rahmenbedingungen des Experimentes (etwa, indem sie den Probandinnen und Probanden kein direktes Feedback zur Richtigkeit der Prognose gaben), schnitten algorithmische Vorhersagesysteme besser als die menschlichen Vorhersagen ab.

35 Dass die Forschungen von Buolamwini/Gebru (2018) von den Anbietern der algorithmischen Klassifikationssysteme zur Kenntnis genommen wurden und dort Änderungen bewirkten, zeigt etwa die entsprechende Stellungnahme von IBM (Puri 2018).

36 Übersetzung TAB; im Original: »Significant subgroup performance disparities persist.«

algorithmische Auswertung visueller Daten in einer Vielzahl von gesellschaftlichen Bereichen (Strafverfolgung, Gesetzesvollzug, innere Sicherheit und Prävention) (Garvie et al. 2016, S. 10 ff.) eingesetzt wird, ist die schlechtere Erkennung von Menschen mit dunklerer Hautfarbe in vieler Hinsicht hoch problematisch.

Häufigere Fehler in der Gesichtserkennung bei Frauen und bei Menschen mit dunkler Hautfarbe sind auch in anderen Einsatzbereichen dokumentiert. So zeigt die unter Mitarbeit des FBI entstandene Studie von Klare et al. (2012), dass sechs getestete algorithmische Gesichtserkennungssysteme Frauen, Personen mit dunkler Hautfarbe und jüngere Menschen zwischen 18 und 30 Jahren deutlich schlechter den passenden Bildern aus der Datenbank zuordnen konnten. Die Trefferquote lag dabei je nach System bei teilweise nur 55 % (Klare et al. 2012, S. 1797).[37] Die Autoren führen diese hohe Fehlerquote auf Trainingsmaterial zurück, dass überdurchschnittlich viele Bilder von hellhäutigen Männern enthält.

Jüngere Studien bestätigen die ungleiche Erkennungsrate bei verschiedenen demografischen Gruppen. Snow (2018) untersuchte im Auftrag der US-amerikanischen Bürgerrechtsorganisation American Civil Liberties Union (ACLU) die Trefferquote des Gesichtserkennungssystems »Rekognition« des Onlinehändlers Amazon. Dieses wandte er auf die 535 Abgeordneten des US-Kongresses an und entdeckte dabei, dass das Gesichtserkennungssystem 28 Kongressmitglieder fälschlicherweise als Kriminelle identifizierte. Die Bilder der Abgeordneten waren mit öffentlich zugänglichen Daten der Strafverfolgungsbehörden abgeglichen worden, sogenannten Mugshots (insgesamt rund 25.000 solcher Fahndungsbilder). Diese Fehlzuordnungen betrafen in besonderem Maße Abgeordnete mit dunkler Hautfarbe (Snow 2018).[38]

Ein Vergleich von 189 kommerziell verfügbaren Gesichtserkennungssystemen durch das US-amerikanische National Institute of Standards and Technology (NIST) bestätigte die Befunde. Im Abgleich von über 18 Mio. Bildern, die 8,5 Mio. Personen abbildeten, zeigte sich eine deutlich erhöhte Fehlerquote bei Menschen aus West- und Ostafrika sowie aus Ostasien (gegenüber Mittel- und Osteuropäerinnen und -europäern). Eine ebenfalls erhöhte Fehlerquote zeigte sich in der Erkennungsrate von Frauen gegenüber Männern sowie von Kindern und älteren Menschen (gegenüber mittelalten Menschen) (Grother et al. 2019).

37 Bei Frauen lag die Trefferquote bei den verschiedenen Systemen bei 55 bis 89 %, bei Männern bei 63 bis 94 %, bei Personen mit dunkler Hautfarbe bei 68 bis 94 % und bei Menschen im Alter zwischen 18 und 30 Jahren bei 62 bis 92 % (Klare et al. 2012, S. 1797) Eine kumulative Trefferquote für junge Frauen mit dunkler Hautfarbe wurde nicht ermittelt. Zu den individuell teils drastischen Folgen solcher falsch-positiver Ergebnisse siehe beispielsweise Snow (2018) und Fry (2019, 191; 231-232).

38 Unter den fälschlicherweise als Kriminelle identifizierten Abgeordneten befanden sich 39 % Afroamerikanerinnen bzw. Afroamerikaner, während nur 20 % der Kongressabgeordneten Afroamerikanerinnen bzw. Afroamerikaner sind (Snow 2018).

In ihrem Bericht unter dem Titel »The Perpetual Line-up«[39] zu »unregulierter polizeilicher Gesichtserkennung in den USA« arbeiteten Garvie et al. (2016) heraus, dass das FBI in 16 US-amerikanischen Bundesstaaten Gesichtserkennungsalgorithmen nutzt, die Bilddateien u.a. mit Fotos aus amtlichen Lichtbilddokumenten wie Pässen, Personalausweisen und Führerscheinen abgleichen. Diese Nutzung von algorithmischen Gesichtserkennungssystemen ist weitgehend unreguliert und betrifft insgesamt 52 bundesstaatliche und lokale Strafverfolgungsbehörden und damit 117 Mio. volljährige US-Amerikanerinnen und Afroamerikaner. Damit werden biometrische Daten, die nicht im Kontext der Strafverfolgung gewonnen wurden, zur polizeilichen Verwendung freigegeben – in den meisten Staaten ohne jegliche externe und teils auch ohne interne Kontrolle (Garvie et al. 2016, S. 1 ff.). Möglichkeiten des Einsatzes umfassen u. a. polizeiliche Kontrollen von Personen, Identitätsfeststellungen bei Verhaftungen, die Identifizierung von Verdächtigen bei polizeilichen Ermittlungen, z. B. anhand von Überwachungskameramaterial, und potenziell zukünftig die Echtzeitidentifizierung bei Videoüberwachung (Garvie et al. 2016, S. 11 f.; Hayward 2019). Die befragten Entwicklerinnen und Entwickler von zwei der führenden Hersteller von Gesichtserkennungssoftware gaben an, ihre Systeme nicht speziell auf möglicherweise verzerrte Fehlerquoten in Abhängigkeit von der Hautfarbe getestet zu haben (Garvie et al. 2016, S. 55).

Die erhöhte Fehlerquote in Abhängigkeit von der Hautfarbe betrifft nicht nur Algorithmen zur Gesichtserkennung, sondern auch die Erkennung von Menschen im Straßenverkehr, wie sie etwa bei (teil)autonom fahrenden Fahrzeuge eingesetzt wird. So zeigen Wilson et al. (2019) in ihrem Vergleich verschiedener visueller Objekterkennungssysteme, dass Menschen mit dunkler Hautfarbe von den getesteten Systemen deutlich schlechter als Fußgängerinnen und Fußgänger erkannt wurden als Menschen helleren Hauttyps. Diese Verzerrung in der Erkennungsrate ließ sich weder durch Variationen in der Tageszeit und damit den Lichtverhältnissen erklären noch durch schlechte Sichtbarkeit aufgrund von verdeckenden Objekten. Dass (teil)autonome Fahrzeuge ein visuelles Signal zweifelsfrei als Fußgängerin bzw. Fußgänger erkennen, ist insbesondere im Einsatz im Straßenverkehr von zentraler Bedeutung. Die durchgehend schlechteren Ergebnisse beim Erkennen von Personen dunkleren Hauttyps legen nahe, dass zukünftige Fehler von autonomen Fahrzeugen vermutlich nicht gleichmäßig auf verschiedene Bevölkerungsgruppen verteilt sein werden (Wilson et al. 2019).

Visuelle Auswertungssysteme weisen offensichtlich häufig eine Ungleichbehandlung von Personen verschiedener Hauttypen auf. Dies geht vermutlich auf eine Reihe von technischen und sozial-organisatorischen Gründen zurück; so

39 Ein Line-up bezeichnet die Aufstellung von Tatverdächtigen gegenüber den Strafermittlungsbehörden und häufig zusätzlich einem Augenzeugen. Der Titel bezieht sich auf die Möglichkeit des permanenten Abgleichs mit polizeilichen Bilddateien und könnte folglich als unaufhörliche Gegenüberstellung übersetzt werden.

sind die ersten Videokameras mit einem besonderen Fokus auf die Darstellung *heller* Hauttypen entwickelt worden, und viele, insbesondere einfachere Sensoren weisen bis heute eine geringere Sensitivität für dunkle Hauttöne auf (Buolamwini 2017; Sandvig et al. 2016). Als Alltagsbeispiel führt Fussell (2017) den Fall an, dass die Sensoren in automatischen Seifenspendern eine dunkelhäutige Hand nicht als solche erkennen und folglich Seife nur bei hellhäutigen Händen ausgeben. Gleichzeitig führen Trainingsdaten, die bestimmte Bevölkerungsgruppen unterdurchschnittlich oder gar nicht abbilden, dazu, dass in Asien entwickelte Gesichtserkennungssysteme durchschnittliche europäische Gesichter schlechter erkennen als asiatische Gesichtstypen (Garvie et al. 2016, S. 53).

4.5 Gemeinsamkeiten und Unterschiede der vier Fallbeispiele

Die vier dargestellten Fallbeispiele weisen eine Reihe von Überschneidungen auf – und unterscheiden sich zugleich in einer Vielzahl von Merkmalen. Die Gegenüberstellung der Fallbeispiele macht übergreifende Fragen und Aspekte des Einsatzes von AES und ML sichtbar.

Gemeinsam ist allen vier Fallbeispielen bzw. Beispielbereichen, dass die algorithmischen Systeme von denen, über die entschieden wurde, nicht freiwillig genutzt wurden (in keinem der beschriebenen Fälle existierte eine Opt-out-Option oder eine wählbare Alternative für die Betroffenen), und dass sie in allen Fällen dazu eingesetzt wurden, *begrenzte Ressourcen effizienter zu verteilen*. Die jeweilig zu verteilenden Ressourcen umfassen die stationäre klinische Behandlung, die Weiterbildungsangebote des AMS, die Unterbringung im geschlossenen Vollzug für Verurteilte und die polizeilichen Kapazitäten zur Personenkontrolle. In allen Fällen wurde die Ungleichbehandlung durch Dritte festgestellt und führte zu einem breiten medialen Echo. Teile der Öffentlichkeit werteten im Zuge dessen die jeweilige Ungleichbehandlung als Diskriminierung – eine Deutung, der teils widersprochen wurde und die in keinem der geschilderten Fälle von einem Gericht als solche bestätigt wurde. Diskriminierung – so die Einschätzung etwa des AMS – sei durch die Entscheidung eines algorithmischen Systems allein nicht gegeben, sondern entstehe erst durch den anschließenden Umgang mit dem Ergebnis des AES (Bachner 2018; Wimmer 2018a). Dieser *große Stellenwert der sozialen Rahmenbedingungen*, in denen ein AES zum Einsatz kommt, gilt für alle Fallbeispiele. Allerdings sind die sozialen Rahmenbedingungen in ihrer möglicherweise diskriminierenden Wirkung schwerer zu bewerten als ein algorithmisches Ergebnis, bei dem Eingabedaten (Input), Datendurchsatz (Throughput) und Ergebnis (Output) nachvollziehbar sind und in klaren kausalen Zusammenhängen stehen.

Denn bei realen Entscheidungssituationen sind zur Beurteilung auf Diskriminierungsrisiken, die durch die Interaktion von AES und menschlichen Entscheidenden entstehen, mehr Faktoren zu berücksichtigen, als dies in einem AES üblicherweise berücksichtigt wird. So dürfte etwa im ersten Fallbeispiel die Entscheidung einer Ärztin darüber, ob ein Patient mit Lungenentzündung stationär oder ambulant behandelt wird, von dem Input abhängen, der auch dem Algorithmus zur Verfügung steht, aber darüber hinaus von anderen Informationen, die aus dem Gespräch mit dem Patienten und ggf. einer vertieften Untersuchung des Patienten stammen. Möglicherweise spielen auch Informationen über krankenhausinterne Parameter (z. B. die derzeitige Belegungssituation einzelner Stationen) und ökonomische Überlegungen (z. B. die Abrechenbarkeit einzelner Leistungen) eine Rolle (Vogd 2004). Der folgende Throughput als Abwägungsprozess zwischen verschiedenen Optionen ist schwieriger nachzuvollziehen als bei einem AES, das in vielen Fällen eindeutigen Berechnungsvorschriften folgt. Wie die Ärztin entscheidet und welche Informationen mit welchem Gewicht in die Abwägung einfließen, ließe sich wohl am besten durch eine diskursive Darlegung nachvollziehen. Der Output schließlich, also die ärztliche Entscheidung über eine stationäre oder ambulante Behandlung, erfolgt in ähnlicher Form (vordergründig) eindeutig wie bei einem AES, ist aber de facto (anders als viele Outputs von vollautomatisierten AES) gegenüber verschiedenen Beteiligten rechtfertigungsbedürftig – z. B. gegenüber dem Patienten, gegenüber Kollegen und Kolleginnen, potenziell auch gegenüber der Geschäftsführung des Krankenhauses und der Krankenversicherung. Der soziale Kontext, in dem ein AES genutzt wird und in dem auf Basis eines empfehlenden AES entschieden und gehandelt wird, ist folglich von zentraler Bedeutung.

Gemeinsamkeiten der dargestellten Fallbeispiele

- Unfreiwilligkeit der Nutzung für von AES Betroffene
- Ziel: effizientere Verteilung begrenzter Ressourcen
- Veröffentlichung der Ungleichheiten durch Dritte, breites mediales Echo, keine gerichtliche Klärung
- Soziale Rahmenbedingungen des Einsatzes werden als zentral für mögliche Diskriminierung angesehen.

Die Fallbeispiele weisen neben den skizzierten Überschneidungen auch eine Reihe von Unterschieden auf, etwa hinsichtlich des *gesellschaftlichen Bereichs*, in dem das jeweilige AES Einsatz findet (Gesundheitsversorgung, Sozialleistungen, Strafvollzug und Strafverfolgung/Straßenverkehr). Verknüpft mit dem gesellschaftlichen Feld des Einsatzes ist auch die *potenzielle Schadenstiefe*, falls ein AES diskriminierende Ergebnisse hervorbringt bzw. Folgen auslöst – diese dürfte bei Algorithmen zur Vorhersage der Sterbewahrscheinlichkeit und damit

verbundener Therapiemaßnahmen mindestens kurzfristig größer sein als bei einer algorithmischen Zuteilung von Qualifikationsmaßnahmen für Arbeitssuchende. Bei diesen beiden Fallbeispielen zeigen sich darüber hinaus zwei weitere Unterscheidungen, die für eine mögliche diskriminierende Wirkung relevant sein können, nämlich die *Einsehbarkeit und Nachvollziehbarkeit der algorithmischen Entscheidungsprozesse* (die im Fall des Algorithmus von Aspire Health eher gering ist, während beim AMS-Algorithmus sogar die Berechnungsvorschriften veröffentlicht wurden) sowie die *Entwickler und Nutzer der algorithmischen Systeme* (öffentliche Hand/Private). Handelt es sich um einen Entwickler von AES, der wirtschaftliche Interessen verfolgt, dürfte das Geschäftsgeheimnis der Veröffentlichung der Berechnungsvorschriften entgegenstehen, wohingegen bei einem nichtkommerziellen Entwickler Transparenz tendenziell leichter herzustellen ist. Wird ein AES durch eine öffentliche Stelle eingesetzt, steht sie potenziell unter größerem Nachweisdruck, keine diskriminierenden Ergebnisse hervorzubringen (etwa, wenn eine fehlerhafte Gesichtserkennung im Bereich der Polizeiarbeit eingesetzt wird), als bei einem privaten Akteur (zum Beispiel zur Identifizierung im vernetzten Zuhause).

Unterschiede der dargestellten Fallbeispiele

- gesellschaftlicher Bereich des Einsatzes
- potenzielle Schadenstiefe einer Diskriminierung
- Einsehbarkeit und Nachvollziehbarkeit der algorithmischen Entscheidungsprozesse
- wirtschaftliches Interesse der Entwickler des AES
- Einsatz der AES durch Private oder die öffentliche Hand
- (vermeintlich) benachteiligte Gruppe
- ökonomisches Risiko für die Algorithmenhersteller im Fall angenommener/nachgewiesener Ungleichbehandlung durch das AES

Eine weitere wichtige Frage, in der sich die vier Fallbeispiele unterscheiden, zielt darauf, wer durch die algorithmische Ungleichbehandlung geschädigt wird. Handelt es sich um eine Gruppe, die nachgewiesenermaßen gesamtgesellschaftlich oder in gesellschaftlichen Teilbereichen bereits benachteiligt ist (im COMPAS-Fall Afroamerikanerinnen und -amerikaner, beim AMS Frauen, insbesondere mit Betreuungspflichten)? Falls ja, dient die algorithmische Ungleichbehandlung dazu, gesellschaftliche Nachteile zu mildern (wie beispielsweise im AMS-Fallbeispiel postuliert)? Im COMPAS-Beispiel etwa führte die Nichtberücksichtigung des geschützten Merkmals Geschlecht dazu, dass Frauen (die insgesamt seltener rückfällig werden) die gleichen Rückfallrisiken attestiert wurden wie Männer. Wann also ist eine Berücksichtigung geschützter Merkmale geboten, wann ist sie diskriminierend? Von Bedeutung ist zudem, ob

die Entwickler eines AES von einem möglicherweise diskriminierenden Output ihres Systems in der Folge marktwirtschaftliche Nachteile zu erwarten haben. Dies ist etwa bei den ungleichen Fehlerraten in der Gesichtserkennung der Fall, wie die umgehende Reaktion von IBM (Puri 2018) auf die Veröffentlichung von Buolamwini/Gebru (2018) zeigt. Im Fall des Aspire-Algorithmus hat demgegenüber auch die öffentliche Berichterstattung nicht zu einer Offenlegung der Berechnungsvorschriften und/oder zu einer Modifikation des AES geführt.

Zusammenfassend zeigt sich, dass zentrale Fragen mit Blick auf AES insbesondere die sozialen Rahmenbedingungen des Einsatzes, die Einseh- und Nachvollziehbarkeit des Algorithmus, die mögliche Schadenstiefe der Differenzierungsentscheidungen sowie die Eigenschaften der vom möglichen Schaden Betroffenen (z. B. bereits benachteiligt) und schließlich die Freiwilligkeit der Nutzung des AES umfassen.

5 Handlungsoptionen

Komplexe AES sind bereits heute Teil der Lebensrealität der allermeisten Menschen, sie »treffen wichtige Differenzierungsentscheidungen in Lebensbereichen, die für Menschen mitunter existenzielle Bedeutung aufweisen« (Martini 2019, S. 333), etwa bei der Kreditvergabe, der Strafverfolgung, der öffentlichen Verwaltung von Sozialleistungen und der Job- und Partnersuche.[40] Es ist anzunehmen, dass diese Systeme in Zukunft aufgrund des technischen Fortschritts und der Effizienz in der Ordnung und Klassifizierung von Informationen eine zunehmend große Rolle spielen werden. Befragungen zeigen, dass die meisten Menschen eher vage Vorstellungen vom Funktionieren und von den Einsatzfeldern und Konsequenzen lernender Systeme haben und dass sie diese zwiespältig betrachten. An dieser Zwiespältigkeit und dem mangelnden Wissen um die Verwendung algorithmischer Systeme knüpft eine Reihe von Forderungen aus dem Forschungsfeld »Algorithmen und Gesellschaft« an, die auf mehr Kontrolle und Transparenz bei der Verwendung komplexer algorithmischer Systeme dringen, denn »ein elaboriertes normatives Algorithmenregulierungsregime für die digitale Entscheidungsunterstützung fehlt bislang« (Martini 2019, S. 340). Tabelle 5.1 gibt einen Überblick über die Ziele (Herstellung von Transparenz, Kontrolle und Evaluierung sowie einheitlicher Regulierung) und Maßnahmen, mittels derer die Ziele erreicht werden sollen.

Eine Reihe von Handlungsoptionen, die in der wissenschaftlichen Debatte zu komplexen algorithmischen Systemen und ihren gesellschaftlichen Folgewirkungen diskutiert werden, zielt auf die Herstellung von *Transparenz*. Transparenz stellt eine zentrale Voraussetzung dafür dar, dass AES darauf überprüfbar sind, ob sie mit legitimen Mitteln einen legitimen Zweck verfolgen. Gegenüber Aufsichtsbehörden und anderen autorisierten Kontrollstellen muss diese Transparenz gegeben sein, um entscheiden zu können, ob die rechtlichen und technischen Vorgaben durch ein AES erfüllt werden (Datenethikkommission 2019, S. 169).

Konkret umfassen die bisher diskutierten Vorschläge u.a. eine Kennzeichnungspflicht, die Nutzenden und Betroffenen überhaupt deutlich macht, dass sie mit einem (lernenden) algorithmischen System interagieren (Beining 2019, S. 11; Martini 2019, S. 341). Diese Kennzeichnungspflicht sollte mit konkreten Umsetzungsvorgaben einhergehen, »um zu verhindern, dass zu allgemeine Auflagen in ›ADM-Erklärungen‹ münden, die zu keinem Wissenszuwachs bei den Betroffenen führen, wie es beispielsweise bei gebräuchlichen Datenschutzerklärungen häufig der Fall ist« (Beining 2019, S. 32). Ebenso sollten den Betroffenen

40 Weitere Details und Handlungsempfehlungen finden sich auch im Bericht der Enquete-Kommission Künstliche Intelligenz – gesellschaftliche Verantwortung und wirtschaftliche, soziale und ökologische Potenziale des Deutschen Bundestages (EK 2020).

von algorithmischen Entscheidungen die Logik und die Gründe für die jeweilige Entscheidung offen gelegt werden (insbesondere in zentralen Lebensbereichen und bei ablehnenden Entscheidungen; Beining 2019, S. 11; Martini 2019, S. 343; Zweig et al. 2018, S. 199). Auch sollten Betroffene mögliche Handlungsoptionen erfahren (etwa hinsichtlich möglicher Korrekturen oder Zweitbeurteilungen; Beck et al. 2019, S. 19; Beining 2019, S. 32; Fry 2019, S. 33).

Tab. 5.1 Handlungsoptionen zur Reduktion von Diskriminierungsrisiken durch komplexe algorithmische Entscheidungssysteme

Herstellung von	Maßnahmen
Transparenz	Kennzeichnungspflicht
	Informationen über Logik von Profilbildungen und Gründe für die jeweilige Entscheidung
	Dokumentation und Protokollierung der Datensätze und Modelle
	Informationszugang für die Allgemeinheit bei Systemen von gesellschaftlichem Interesse
Kontrolle und Evaluierung	Schaffung eines bundesweiten Kompetenzzentrums zu algorithmischen Systemen
	Etablierung eines risikoadaptierten Regulierungsansatzes mit Einteilung in verschiedene Kritikalitätsstufen
	Einführung eines Zulassungssystems zur Vorabprüfung
	Analyse der Ergebnisse von algorithmischen Entscheidungsverfahren durch Antidiskriminierungsstellen
	Risikofolgenabschätzung durch Betreiber sowie Verpflichtung zu mathematisch-prozeduralen Qualitätsgarantien
einheitlicher Regulierung	horizontale Vorgaben im Recht der EU
	Ausweitung des Schutzes vor Diskriminierung anhand von Gruppenmerkmalen
	Prüfung und Anpassung des Haftungsrechtes
	Einführung eines kollektiven Rechtsschutzes mit Möglichkeit der Verbandsklage
	regulierte Selbstregulierung (Zertifizierung/Codex)

Eigene Darstellung auf Basis von Beining 2019, S. 31 ff.; Datenethikkommission 2019; Martini 2019, S. 333 ff.; Orwat 2020, S. 97 ff.; Saurwein 2019; Zweig/Krafft 2018; Zweig et al. 2018

Gerade im Hinblick auf mögliche Beschwerden von Nutzenden erscheint es sinnvoll, die Betreiber von AES zu einer Protokollierung der verwendeten Datensätze und Modelle zu verpflichten, um auch ex post mögliche Fälle von Diskriminierung nachvollziehen zu können (Datenethikkommission 2019, S. 22; Martini 2019, S. 353 f.; Zweig et al. 2018, S. 198). Hierzu könnten auch Ergänzungen und Konkretisierungen der bestehenden Vorgaben der Datenschutzgrundverordnung zu Informationspflichten, Auskunftsrechten und Vorgaben zu Dokumentationen erwogen werden (Orwat 2020, S. 106 ff. u 119 ff.).

Grundsätzlich ist bei Regulierungsinstrumenten, die Informationspflichten oder Auskunftsansprüche begründen, der Schutz von Betriebs- und Geschäftsgeheimnissen zu berücksichtigen. Doch müssen diese bei einer Offenlegung von Kriterien, wie eine Entscheidung zustande gekommen ist, und der Gewichtungen der Kriterien untereinander nicht notwendigerweise preisgegeben werden, da nicht das komplette algorithmische Verfahren vollständig offengelegt werden muss. Ferner muss, wenn automatisierte Entscheidungen im Einzelfall im Sinne des Artikels 22 Datenschutz-Grundverordnung vorliegen, ohnehin nach Artikle 14 Datenschutz-Grundverordnung über die involvierte Logik, deren Tragweite und angestrebten Auswirkungen der Datenverarbeitung informiert werden. Hier hat der Gesetzgeber den berechtigten Informationsinteressen der Betroffenen einen hohen Stellenwert eingeräumt.

Zudem könnte der Schutz von Betriebs- und Geschäftsgeheimnissen auch dadurch sichergestellt werden, indem staatliche Aufsichtseinrichtungen, wie z. B. die Aufsichtsbehörden des Datenschutzes oder Antidiskriminierungsstellen, oder vertrauenswürdige (private) Dritte bzw. Auditoren, wie man diese etwa aus dem Bereich der Wirtschaftsprüfung kennt, zwischengeschaltet werden. Diese könnten die notwendigen Informationen so verdichten und aufbereiten, dass keine Betriebs- und Geschäftsgeheimnisse offengelegt werden. Des Weiteren bestünde die Möglichkeit, dass sich die Pflichten zur Information, Dokumentation und Zurverfügungstellung nur auf eine Informierung der Aufsichtseinrichtungen bezieht, die dann die Schutzinteressen der Berechtigten durchsetzen.

Mit Blick auf Systeme von gesellschaftlichem Interesse schlägt die Datenethikkommission (2019, S. 22) vor, einen Informationszugang für die Allgemeinheit rechtlich zu verankern, der es beispielsweise Journalisten oder Forschenden ermöglicht, potenziell diskriminierende Verfahren im Detail reproduzieren zu können.

Befragt nach den Maßnahmen, die sie sich mit Blick auf komplexe AES wünschen, sprechen sich in einer repräsentativen Bevölkerungsbefragung die meisten Bürgerinnen und Bürger für solche Maßnahmen aus, die auf die Herstellung von Transparenz zielen. So stimmten die meisten Befragten für ein Recht auf eine zweite Beurteilung (81 %), ein Auskunftsrecht für Betroffene

(80 %) sowie eine Kennzeichnungspflicht bei Entscheidungen durch Algorithmen (79 %) (Fischer/Petersen 2018, S. 29). Zu wissen, dass algorithmische Systeme zum Einsatz kommen, und diese ggf. durch eine zweite Beurteilung validieren zu lassen, scheint für die meisten Befragten von zentraler Bedeutung.

An dritter und vierter Stelle der am häufigsten genannten Forderungen mit Blick auf Algorithmen nannten die repräsentativ Befragten Maßnahmen, die sich unter das Ziel der *Kontrolle und Evaluierung* einordnen lassen. So fordern jeweils drei von vier Befragten eine Einsicht und Prüfung von Algorithmen durch unabhängige Experten (75 %) sowie die Einführung eines TÜV-Äquivalents für Algorithmen (74 %) (Fischer/Petersen 2018, S. 29). Der Forderung nach einem Algorithmen-TÜV bzw. einer bundeseinheitlichen Prüfstelle für algorithmischen Systemen schließt sich eine Reihe von Autoren an (Beck et al. 2019, S. 16; Datenethikkommission 2019, S. 23; Zweig et al. 2018, S. 199).[41] »Genau wie es die US Food and Drug Administration bei Arzneimitteln macht, würde diese Behörde hinter verschlossenen Türen Algorithmen auf Genauigkeit, Einheitlichkeit und Verzerrungen testen, und sie hätte die Befugnis, die Anwendung eines Produkts bei echten Menschen zu genehmigen oder zu verweigern« (Fry 2019, S. 88). Allerdings wendet Martini (2019, S. 354) ein, dass eine einheitliche Bundeskontrollstelle für lernende Systeme bzw. komplexe algorithmische Systeme zwar wünschenswert, aufgrund der föderalen Struktur Deutschlands jedoch schwerlich realisierbar sei. Denkbar sei hingegen eine bundesbehördliche Organisation, die bestehende Bundes- und Landesbehörden bei der Aufsicht unterstützt.

Ferner sollte die Stärkung der Kompetenzen und Ressourcen der Antidiskriminierungstellen erwogen werden, vor allem, weil es für die Betroffenen erschwert oder sogar unmöglich ist, eine durch ein AES verursachte Diskriminierung zu erkennen und notwendige erste Indizien beizubringen. Denn nach dem AGG muss eine Person, die behauptet, diskriminiert worden zu sein, zunächst mit Indizien nachweisen, dass sie anders als andere Personen behandelt wurde, dass sie sich im Hinblick auf eines der geschützten Merkmale unterscheidet und dass das geschützte Merkmal ursächlich für die Diskriminierung war. Dieser Nachweis mit Indizien ist bei algorithmenbasierten Entscheidungen, die stark personalisiert sind, die sich dynamisch ändern können und die vor allem auf Basis von Ersatzmerkmalen bzw. Proxies, die mit den geschützten Merkmalen korrelieren, erfolgen können, schwer oder überhaupt nicht zu erbringen. Daher sollten Antidiskriminierungsstellen quasi als Stellvertreter stärker selbstständig investigativ vorgehen und rechtliche Schritte einleiten können. Dies erscheint

41 Im Frühjahr 2020 ist ein deutsches KI-Observatorium beim Bundesministerium für Arbeit und Soziales (BMAS 2019) eingerichtet worden, dessen Kernaufgaben »die Beobachtung von Technologieentwicklung, -verbreitung und Technologiefolgenabschätzung für KI in Arbeit und Gesellschaft« umfassen sollen. Auf längere Sicht soll ein eigenes Bundesinstitut für künstliche Intelligenz eingerichtet werden, das bei der Bewertung der Technologie und der politischen Steuerung mitwirken soll (Krempl 2019).

insbesondere hinsichtlich ihres gesetzlichen Schutzauftrags notwendig. Des Weiteren können sie beratend bei der Entwicklung und Anwendung von AES tätig werden und zur Diskriminierungsprävention beitragen (Orwat 2020, S. 129 ff.).

Zugleich gilt zu bedenken, dass insbesondere lernende Systeme sich beständig verändern und deshalb eine Ex-ante-Evaluierung vor ihrem Einsatz kein vollständiges, dauerhaft gültiges Bild ergeben kann. Oftmals liegen die Risiken zudem nicht so sehr in den Algorithmen selbst, sondern in den Geschäftsmodellen (und Entscheidungspraxen), die auf diesen algorithmischen Systemen fußen (Saurwein 2019, S. 49). Generell müsse man, um die Folgen eines Algorithmus abschätzen zu können, diesen in den Situationen und sozialen Handlungskontexten evaluieren, in denen er eingesetzt wird. Eine Beurteilung eines Algorithmus fernab seines Einsatzkontextes habe nur wenig Aussagekraft über gesellschaftlichen Wirkungen und erlaube keine abschließende ethische Einschätzung, so Sandvig et al. (2016, S. 4982).

Einen strukturellen Rahmen zur Beurteilung algorithmischer Systeme schlägt die Datenethikkommission (2019, S. 163 ff.) mit ihrem Entwurf eines risikoadaptierten Regulierungsansatzes mit Einteilung in verschiedene Kritikalitätsstufen vor. Dabei soll sich der Grad der Regulierung, der ein algorithmisches System unterworfen wird, an der jeweiligen *Systemkritikalität* orientieren: »Die Systemkritikalität setzt am Schädigungspotenzial des algorithmischen Systems an. Dabei bedeutet Schädigungspotenzial die Kombination aus der Wahrscheinlichkeit des Schadenseintritts und der Schwere des zu befürchtenden Schadens.« (Datenethikkommission 2019, S. 18) Beschreibung und Definition, was genau als Schaden bzw. Schädigungspotential zu verstehen ist und welche Risikozusammenhänge für Grundrechte bestehen, werden allerdings von der Datenethikkommission nicht geleistet. Hier könnten sich Herausforderungen bei der Bildung und Abgrenzung von Risikoklassen und der Zuordnung von (soziotechnischen) Systemen zu ihnen ergeben.

Die in Abhängigkeit von der Systemkritikalität vorzunehmenden Regulierungsmaßnahmen reichen von einem vollständigen oder teilweisen Verbot über Kontroll- und Transparenzpflichten bis hin zu Auditverfahren. Nicht schädigungsfähige Anwendungen erfordern keine gesonderten Maßnahmen (Datenethikkommission 2019, S. 20 ff.).[42] Dies ließe sich zu einem staatlichen Zulassungssystem für komplexe AES erweitern. Bei Systemen mit erheblichem Schädigungspotenzial fordert die Datenethikkommission (2019, S. 20) umfassende Offenlegungspflichten hinsichtlich der zugrundeliegenden Berechnungen und Daten und die Möglichkeit für Kontrollbehörden, über eine Liveschnittstelle in

42 Martini (2019, S. 350) schlägt ein ähnliches Prüfschema vor, das sich auch auf die Anwender bezieht und für öffentliche Stellen eine strengere Prüfung vorsieht als für nichtöffentliche Stellen.

Echtzeit das System überprüfen zu können. Für Anwendungen mit unvertretbarem Schädigungspotenzial müsste ein partielles oder komplettes Verbot ausgesprochen werden. Zugleich sollte bei der Prüfung in Betracht gezogen werden, inwiefern ein System umgehbar ist bzw. ob es Alternativen zur Nutzung des jeweiligen Systems gibt: »Je monopolistischer ein algorithmisches Entscheidungssystem verwendet wird, desto besser muss die Qualitätssicherung sein.« (Zweig/Krafft 2018, S. 220)

Ein weiterer Vorschlag zur Kontrolle und Evaluierung komplexer AES zielt darauf ab, eine Risikofolgenabschätzung durch die Betreiber sowie eine Verpflichtung zu mathematisch-prozeduralen Qualitätsgarantien zu etablieren (Martini 2019, S. 346 ff.). Die Datenschutz-Grundverordnung sieht nach Artikel 35, Absatz 1, Satz 1 bereits heute bei absehbar hohem Risiko für die Rechte und Freiheiten natürlicher Personen eine Folgeabschätzung für algorithmenbasierte Anwendungen vor. Diese Verpflichtung zur Folgeabschätzung sollte nach Ansicht von Martini (2019, S. 346) auch auf mögliche Folgen für Rechte und Interessen der Betroffenen (etwa hinsichtlich Diskriminierungsrisiken) ausgedehnt werden.[43] Je nach Kritikalität der Anwendung könnten Hersteller und Betreiber komplexer algorithmischer Systeme gestuften Auskunfts-, Mitwirkung- und Berichtspflichten unterworfen werden; ebenso könnten Trainingsprozesse standardisiert erfolgen und die mathematisch-statistische Validität der Ergebnisse von AES kontrolliert werden. Auch die Pflicht zur Einbindung von Risikomanagementsystemen in algorithmische Anwendung wäre denkbar, mit dem Ziel, dass diese etwa Anzeichen mittelbarer Diskriminierungen erkennen und eine Kontrolle der automatisierten Entscheidung erwirken (Martini 2019, S. 352 f.). »Kontrollalgorithmen […] richten damit die Künstliche Intelligenz gleichsam gegen sich selbst: Sie analysieren, welche Faktoren die implementierten Algorithmen besonders stark gewichten und ob die nach außen kommunizierte Entscheidungslogik mit dem tatsächlichen Entscheidungsverhalten übereinstimmt.« (Martini 2019, S. 351)

Um EU-weit eine möglichst *einheitliche Regulierung* zu fördern, sollten die skizzierten Maßnahmen (insbesondere zur Einordnung der Systemkritikalität, zu den Rechten von Betroffenen sowie zu Kontrollinstitutionen und Strukturen) laut Datenethikkommission (2019, S. 180 ff.) auf europäischer Ebene in die Gestaltung einer horizontalen Regelung zu algorithmischen Systemen einfließen. In einer solchen zu formulierenden europäischen Verordnung für Algorithmische Systeme (EUVAS) sollten zentrale Grundprinzipien für den Einsatz und die Gestaltung der Systeme festgelegt werden, um so den Bürgerinnen bzw. Bürgern der EU Erwartungssicherheit zu geben.

43 Allerdings sieht die Datenschutzkonferenz (2018a u. 2018b) in ihren Handreichungen für die Datenschutz-Folgenabschätzung und für die Bestimmung von Risiken nach der Datenschutz-Grundverordnung bereits die Abschätzung auch auf mögliche Diskriminierungen und weitere Implikationen für die Grundrechte und Freiheiten der Bürgerinnen und Bürger vor.

Zu erwägen ist darüber hinaus nach Ansicht sowohl von Martini (2019, S. 349) als auch der Datenethikkommission (2019, S. 194) eine Ausweitung des Diskriminierungsschutzes, wie ihn das AGG vorsieht. Konkret schlägt Martini (2019, S. 349) zwei Möglichkeiten vor, nämlich entweder die Ausweitung des AGG »auf alle algorithmenbasierte Verfahren, die einen besonderen Gefährdungsgrad für Persönlichkeitsrechte aufweisen« oder die Erweiterung des AGG um »einzelne Sachbereiche besonders persönlichkeitssensitiver Entscheidungen« (wie etwa Kreditscoring oder Gesichtserkennung).

Zudem könne nach Martini (2019, S. 247 f. u. 349) die Beweislastumkehr, die das AGG für Diskriminierungsfälle bereits vorsieht, so erweitert werden, dass das Vorliegen einer Blackboxauswertung als Indiz für mögliche Diskriminierungen bereits ausreiche und somit bei jeder Blackboxauswertung die Beweislastumkehr gelte. Eine Öffnung der bislang besonders geschützten Merkmale würde den durch Algorithmen zu erwartenden Ungleichbehandlungen (beispielsweise anhand des Wohnortes) und den damit verbundenen »ganz neuen Gerechtigkeitsfragen« potenziell eher gerecht werden als die derzeitige Regelung (Datenethikkommission 2019, S. 194; siehe auch Orwat 2020, S. 112). Verstöße gegen das AGG könnten zusätzlich zu einer möglichen Schadensersatzpflicht auch mit finanziellen Sanktionen bewehrt werden (Martini 2019, S. 356).

In diesem Sinne sollte auch das Haftungsrecht geprüft und ggf. angepasst werden. Derzeit bestehen nach Einschätzung der Datenethikkommission (2019, S. 220 ff.) Unsicherheiten und Haftungslücken beim Einsatz komplexer algorithmischer Systeme, die sich als hinderlich für die Förderung und Akzeptanz neuer Innovationen erweisen könnten. Deshalb wäre eine Anpassung des Produkthaftungsrechtes und ggf. der Verschuldens- und Gefährdungshaftung im Sinne der Rechtsklarheit und Rechtssicherheit angezeigt (eine ähnliche Forderung findet sich auch bei Beck et al. 2019, S. 17). Neben dem Haftungsrecht sollte nach Ansicht einiger Autoren und Autorinnen auch die Verbandsklage reformiert werden. Derzeit haben Verbraucherverbände nach dem Unterlassungsklagengesetz (UklaG, § 3) bei Datenschutzverstößen ein Klagerecht, dieses ließe sich für gemeinnützige Vereinigungen, insbesondere Antidiskriminierungsstellen, auf AES ausdehnen (Martini 2019, S. 357; in ähnlicher Form Datenethikkommission 2019, S. 204; Orwat 2020, S. 108 f.).

Schließlich könnte eine regulierte Selbstregulierung mittels Zertifizierungen und/oder Codizes zum Einsatz kommen. Betreiber und Hersteller könnten an eine Selbstverpflichtung mit bestimmten inhaltlichen Mindestanforderungen gebunden werden, die Veröffentlichungspflichten und Sanktionen bei nichteingehaltenen Versprechen umfasst. Ein solcher branchenspezifischer Codex könnte bei Nichtbeachtung neben wirtschaftlich relevantem Reputationsverlust zudem etwa Bußgeldsanktionen nach sich ziehen (Martini 2019, S. 359; in ähnlicher Form Beining 2019; Datenethikkommission 2019, S. 23 f.).

Neben diesen konkreten Maßnahmen, die in erster Linie auf eine Veränderung regulatorischer Vorgaben und technischer Gestaltung zielen, wird eine Herangehensweisen vorgeschlagen, die sich auf die gesamte Gesellschaft oder auf einzelne gesellschaftliche Teilgruppen richten. So sieht etwa Hagendorff (2019b, S. 131) es mit Blick auf die Fortsetzung und Potenzierung menschlicher Vorurteile in algorithmischen Systemen als zentral an, das »ideologische Setting« der Gesellschaft zu verändern, um Diskriminierungen gleichsam an ihren Ursprüngen zu verhindern. In eine ähnliche Richtung zielen Ansätze, die einen breiten gesellschaftlichen Diskurs über komplexe AES, ihre Wertsetzungen und gesellschaftlichen Folgewirkungen anstoßen wollen: »Welche Kriterien für Nichtdiskriminierung und Gerechtigkeit in welchem Kontext angemessen sind, ist keine technische, sondern eine gesellschaftliche und politische Frage. Daher dürfen diese Fragen auch nicht allein den Technik-Entwicklern überlassen werden.« (Datenethikkommission 2019, S. 169) Teil dieser Diskussion müsste auch sein, welche Grenzen AES gesetzt werden, in welchen Kontexten diese wie eingesetzt werden können und sollen und »ob es verallgemeinerte Situationen gibt, in denen algorithmische Entscheidungssysteme gar nicht eingesetzt werden sollten« (Zweig/Krafft 2018, S. 219).

Im Zuge dieses gesellschaftlichen Diskurses sollte die »algorithmic literacy« in der Bevölkerung gefördert werden, um die individuelle Auseinandersetzung mit komplexen algorithmischen Systemen zu ermöglichen (Beining 2019, S. 33). Umgekehrt sollte die soziale, ethische und politische Verantwortung, die mit der Entwicklung von komplexen algorithmischen Systemen einhergeht, (angehenden) IT-Spezialistinnen und -Spezialisten bewusst sein: »Verantwortliche in der Wissenschaft und der Bildungspolitik sollten dafür Sorge tragen, dass ADM-Entwickler[innen und -Entwickler] von morgen die notwendigen Kompetenzen im Rahmen aktualisierter Curricula, neuer sozioinformatischer Studiengänge und durch die Zusammenarbeit in interdisziplinären Teams erwerben können.« (Beining 2019, S. 33)

Mit Blick auf die Forschungsförderung sehen Krafft/Zweig (2018, S. 488) die Notwendigkeit, einen standardisierten Entwicklungs- und Evaluationsprozess zu erarbeiten, mit dem AES in der öffentlichen IT (wie z. B. im Justizsystem) bewertet werden. Auch ein anderer Bereich der Kontrolle lernender Systeme bedarf weiterer Forschung, nämlich die »interpretable/explainable AI« – lernende Systeme, deren Funktionsweise eine Transparenz des Entscheidungsprozesses sowie eine Begründung getroffener Entscheidung beinhaltet (Datenethikkommission 2019, S. 21). Insgesamt bestehen hinsichtlich der Interaktion zwischen Menschen und lernenden Systemen vielfältige Forschungsdesiderate.

Es ist davon auszugehen, dass komplexe AES in zunehmendem Maße Teile gesellschaftlicher und individueller Routinen und damit als zunehmend selbstverständlich angesehen werden. Die in diesem Hintergrundpapier dargestellten Fallbeispiele machen deutlich, dass damit für gesellschaftliche Gruppen zum Teil

zunächst von außen nichtwahrnehmbare Ungleichbehandlungen einhergehen können – möglicherweise, ohne dass das Individuum sich als Teil einer Gruppe versteht und ohne dass dem Individuum deutlich wird, dass eine Ungleichbehandlung vorliegt bzw. woher diese rührt. »In einer digitalen Gesellschaft wird es […] immer häufiger zur Alltagsrealität gehören, dass Individuen kraft algorithmischer Differenzierung eine ungleiche Behandlung erfahren – nicht, weil sie bestimmte Merkmale *realiter* erfüllen, sondern weil ein Algorithmus ihnen diese Merkmale auf der Grundlage einer Gruppenzuordnung zuschreibt.« (Martini 2019, S. 236) Die Entwicklung hin zu einer *algorithmisierten Gesellschaft* birgt Chancen und erfordert zugleich die Reflexion darüber, welche Werte vermittelt über die Gesellschaft in AES walten, woher diese herrühren, welche Folgen sie bei einem breiten gesellschaftlichen Einsatz zeitigen und welche gesellschaftlichen und politischen Regulative sie ggf. erfordern. Wie gezeigt wurde, können weitreichende gesellschaftliche Weichenstellungen bereits bei Entwicklung und Anwendung der Systeme erfolgen, die Konsequenzen auf die Verwirklichung von Gerechtigkeit und Gleichbehandlung, von freier Entfaltung der Persönlichkeit oder der informationellen Selbstbestimmung haben. Diese Weichenstellungen sind jedoch als gesellschaftliche und politische Aufgaben zu behandeln und den legitimierten Verfahren der gesellschaftlichen Meinungsbildung, der demokratischen Prozesse, des Gesetzgebers oder der Rechtsprechung und der demokratischen Kontrolle zu überlassen und zu verantworten.

6 Literatur

Allhutter, D.; Cech, F.; Fischer, F.; Grill, G.; Mager, A. (2020): Algorithmic Profiling of Job Seekers in Austria: How Austerity Politics Are Made Effective. In: Frontiers in Big Data 3(5), doi: 10.3389/fdata.2020.00005

AMS (2019b): Arbeitsmarktdaten-Online. Dezember 2019. Wien, www.ams.at/content/dam/download/allgemeine-informationen/001_Verzeichnis%20Arbeitsmarktbezirke-Internet_Stand_1219.pdf (19.11.2020)

AMS (Arbeitsmarktservice Österreich) (2019a): Arbeitsmarkt in Karten. Dezember 2019. Wien, www.ams.at/content/dam/download/arbeitsmarktdaten/%C3%B6sterreich/berichte-auswertungen/001_am_karten_1219.pdf (19.11.2020)

Angwin, J.; Larson, J.; Mattu, S.; Kichner, L. (2016): Machine Bias. ProPublica, 23.5.2016, www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (19.11.2020)

Avati, A.; Jung, K.; Harman, S.; Downing, L.; Ng, A.; Shah, N. (2018): Improving palliative care with deep learning. In: BMC Medical Informatics and Decision Making 18(Suppl 4), doi: 10.1186/s12911-018-0677-8

Bachner, M. (2018): AMS-Chef Kopf: »Algorithmus ist Vorteil für Frauen«. In: Kurier, 7.12.2018, kurier.at/politik/inland/ams-chef-kopf-algorithmus-ist-vorteil-fuer-frauen/400346614 (19.11.2020)

Barocas, S.; Selbst, A. (2016): Big Data's Disparate Impact. In: California Law Review 104(3), S. 671–732

Beck, S.; Grunwald, A.; Jacob, K.; Matzner, T. (2019): Künstliche Intelligenz und Diskriminierung: Herausforderungen und Lösungsansätze. Whitepaper, Lernende Systeme – Die Plattform für Künstliche Intelligenz (Hg.), München, www.plattform-lernende-systeme.de/publikationen-details/kuenstliche-intelligenz-und-diskriminierung-herausforderungen-und-loesungsansaetze.html (19.11.2020)

Beining, L. (2019): Wie Algorithmen verständlich werden. Ideen für Nachvollziehbarkeit von algorithmischen Entscheidungsprozessen für Betroffene. Bertelsmann Stiftung (Hg.), Berlin

Benjamin, R. (2019): Assessing risk, automating racism. In: Science 366(6464), S. 421–422

Berg, A. (2018): Künstliche Intelligenz. Von der Strategie zum Handeln. Bitcom, Berlin, www.bitkom.org/sites/default/files/2018-12/Bitkom%20Charts%20K%C3%BCnstliche%20Intelligenz%2005%2012%202018_final.pdf (19.11.2020)

Berres, I. (2017): Afroamerikaner sterben im Schnitt vier Jahre früher als Weiße. In: Spiegel, 3.5.2017, www.spiegel.de/gesundheit/diagnose/usa-afroamerikaner-sterben-im-schnitt-vier-jahre-frueher-als-weisse-a-1145880.html (19.11.2020)

Blunden, M. (2020): Booker beware: Airbnb can scan your online life to see if you're a suitable guest. In: Evening Standard, 3.1.2020, www.standard.co.uk/tech/airbnb-software-scan-online-life-suitable-guest-a4325551.html (19.11.2020)

BMAS (Bundesministerium für Arbeit und Soziales) (2019): Ein Jahr Strategie Künstliche Intelligenz der Bundesregierung. Zwischenbericht zur KI Strategie. Pressemitteilung vom 15.11.2019, www.bmas.de/DE/Presse/Pressemitteilungen/2019/ein-jahr-ki-strategie.html (19.11.2020)

Brennan, T.; Dieterich, W.; Ehret, B. (2009): Evaluating the predictive validity of the COMPAS Risk and Needs Assessment System. In: Criminal Justice and Behavior 36(1), S. 21–40

Briseño, C. (2018): Optimierung der Palliativversorgung: Wenn Algorithmen den Tod vorhersagen. In: Algorithmenethik, 21.3.2018, https://algorithmenethik.de/2018/03/21/optimierung-der-palliativversorgung-wenn-algorithmen-den-tod-vorhersagen/ (19.11.2020)

Britz, G. (2007): Freie Entfaltung durch Selbstdarstellung. Eine Rekonstruktion des allgemeinen Persönlichkeitsrechts aus Art. 2 I GG. Tübingen

Britz, G. (2008): Einzelfallgerechtigkeit versus Generalisierung. Verfassungsrechtliche Grenzen statistischer Diskriminierung. Tübingen

Bucher, T. (2017): The algorithmic imaginary: exploring the ordinary affects of Facebook algorithms. In: Information, Communication & Society 20(1), S. 30–44

Bundesregierung (2018): Einsatz von Algorithmen und automatisierten Prozessen in Jobcentern. Antwort auf die Kleine Anfrage der Abgeordneten Jessica Tatti, Susanne Ferschl, Doris Achtelwilm, weiterer Abgeordneter und der Fraktion DIE LINKE. – Drucksache 19/4450 –. Drucksache 19/5014, Berlin

Bundesregierung (2019): Bundesregierung stärkt die Förderung Künstlicher Intelligenz mit zusätzlichen 500 Millionen Euro. Pressemitteilung vom 23.5.2019, www.bmwi.de/Redaktion/DE/Pressemitteilungen/2019/20190523-bundesregierung-staerkt-die-foerderung-kuenstlicher-intelligenz.html (19.11.2020)

Buolamwini, J. (2017): Algorithms aren't racist. Your skin is just too dark. In: Hackernoon, 29.5.2017, hackernoon.com/algorithms-arent-racist-your-skin-is-just-too-dark-4ed31a7304b8 (19.11.2020)

Buolamwini, J. (2018): When the Robot Doesn't See Dark Skin. In: The New York Times, 21.6.2018, www.nytimes.com/2018/06/21/opinion/facial-analysis-technology-bias.html (19.11.2020)

Buolamwini, J.; Gebru, T. (2018): Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In: Proceedings of Machine Learning Research 81, S. 1–15

Busch, C. (2018): Algorithmic Accountability. Universität Osnabrück, Osnabrück, www.abida.de/sites/default/files/ABIDA%20Gutachten%20Algorithmic%20Accountability.pdf (19.11.2020)

Cabitza, F.; Rasoini, R.; Gensini, G. (2017): Unintended Consequences of Machine Learning in Medicine. In: JAMA 318(6), S. 517–518

Caruana, R.; Lou, Y.; Gehrke, J.; Koch, P.; Sturm, M.; Elhadad, N. (2015): Intelligible Models for HealthCare. In: Cao, L.; Zhang, C.; Joachims, T.; Webb, G.; Margineantu, D.; Williams, G. (Hg.): KDD '15: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, S. 1721–1730

Castelluccia, C.; Le Métayer, D. (2019): Understanding algorithmic decision-making: Opportunities and challenges. European Union (Hg.), Brüssel

Cave, S.; Dihal, K. (2019): Hopes and fears for intelligent machines in fiction and reality. In: Nature Machine Intelligence 1(2), S. 74–78

Corbett-Davies, S.; Pierson, E.; Feller, A.; Goel, S. (2016): A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear. In: The Washington Post, 17.10.2016, www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/?noredirect=on (19.11.2020)

Dastin, J. (2018): Amazon scraps secret AI recruiting tool that showed bias against women. In: Reuters, 11.10.2018, www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G (19.11.2020)

Datenethikkommission der Bundesregierung(2019): Gutachten der Datenethikkommission. Berlin, https://datenethikkommission.de/wp-content/uploads/191128_DEK_Gutachten_bf_b.pdf (19.11.2020)

Datenschutzkonferenz (2018a): Datenschutz-Folgenabschätzung nach Art. 35 DS-GVO. Kurzpapier Nr. 5. https://www.datenschutzkonferenz-online.de/media/kp/dsk_kpnr_5.pdf (18.11.2020)

Datenschutzkonferenz (2018b): Risiko für die Rechte und Freiheiten natürlicher Personen. Kurzpapier Nr. 18. https://www.datenschutzkonferenz-online.de/media/kp/dsk_kpnr_18.pdf (18.11.2020)

Desmarais, S.; Singh, J. (2013): Risk Assessment Instruments Validated and Implemented in Correctional Settings in the United States. https://csgjusticecenter.org/wp-content/uploads/2020/02/Risk-Assessment-Instruments-Validated-and-Implemented-in-Correctional-Settings-in-the-United-States.pdf (19.11.2020)

Dickson, E. (2020): Who's Allowed To Use Airbnb? In: Rolling Stone, 8.1.2020, www.rollingstone.com/culture/culture-news/airbnb-sex-worker-discrimination-935048/ (19.11.2020)

Dietvorst, B.; Simmons, J.; Massey, C. (2015): Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err. In: Journal of experimental psychology. General 144(1), S. 114–126

Dressel, J.; Farid, H. (2018): The accuracy, fairness, and limits of predicting recidivism. In: Science advances 4(1), doi: 10.1126/sciadv.aao5580

Dutton, E.; Van der Linden, D.; Madison, G.; Antfolk, J.; Woodley of Menie, M. (2016): The intelligence and personality of Finland's Swedish-speaking minority. In: Personality and Individual Differences 97, S. 45–49

EK (Enquete-Kommission – Gesellschaftliche Verantwortung und wirtschaftliche, soziale und ökologische Potenziale (2020): Bericht der Enquete-Kommission Künstliche Intelligenz –Gesellschaftliche Verantwortung und wirtschaftliche, soziale und ökologische Potenziale. Unterrichtung der Enquete-Kommission Künstliche Intelligenz –Gesellschaftliche Verantwortung und wirtschaftliche, soziale und ökologische Potenziale. Deutscher Bundestag, Drucksache 19/23700, Berlin

Endres, A. (2014): Der Rassismus ist messbar. Schwarze in den USA. In: Zeit Online, 25.8.2014, www.zeit.de/wirtschaft/2014-08/schwarze-usa-soziale-ungleichheit/komplettansicht (19.11.2020)

Ernst, C. (2017): Algorithmische Entscheidungsfindung und personenbezogene Daten. In: Juristenzeitung 72(21), S. 1026–1036

Europäische Kommission (2020): Weißbuch zur Künstlichen Intelligenz – ein europäisches Konzept für Exzellenz und Vertrauen. COM(2020) 65 final, Brüssel, ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_de.pdf (19.11.2020)

Fanta, A. (2018): Österreichs Jobcenter richten künftig mit Hilfe von Software über Arbeitslose. In: Netzpolitik.org, 13.10.2018, netzpolitik.org/2018/oesterreichs-jobcenter-richten-kuenftig-mit-hilfe-von-software-ueber-arbeitslose/ (19.11.2020)

Fischer, S.; Petersen, T. (2018): Was Deutschland über Algorithmen weiß und denkt. Ergebnisse einer repräsentativen Bevölkerungsumfrage. Bertelsmann Stiftung (Hg.), Impuls Algorithmenethik #7, Gütersloh

Fröhlich, W.; Spiecker genannt Döhmann, I. (2019): Können Algorithmen diskriminieren? In: Zeitschrift des Deutschen Juristinnenbundes 22(2), S. 91–93

Fry, H. (2019): Hello World. Was Algorithmen können und wie sie unser Leben verändern. München

Fussell, S. (2017): Why Can't This Soap Dispenser Identify Dark Skin? In: Gizmondo, 17.8.2017, gizmodo.com/why-cant-this-soap-dispenser-identify-dark-skin-1797931773 (19.11.2020)

Gandy Jr., O. (2010): Engaging rational discrimination: exploring reasons for placing regulatory constraints on decision support systems. In: Ethics and Information Technology 12(1), S. 29–42

Garvie, C.; Bedoya, A.; Frankle, J. (2016): The perpetual line-up. Unregulated Police face recognition in America. Georgetown Law, Center on Privacy & Technology (Hg.), https://www.perpetuallineup.org/report (19.11.2020)

Gerberding, J.; Wagner, G. (2019): Qualitätssicherung für »Predictive Analytics« durch digitale Algorithmen. In: Zeitschrift für Rechtspolitik 4, S. 116–118

Grother, P.; Ngan, M.; Hanaoka, K. (2019): Face Recognition Vendor Test (FRVT). Part 3: Demographic Effects. National Institute of Standards and Technology, nvl pubs.nist.gov/nistpubs/ir/2019/NIST.IR.8280.pdf (19.11.2020)

Grzymek, V.; Puntschuh, M. (2019): Was Europa über Algorithmen weiß und denkt. Ergebnisse einer repräsentativen Bevölkerungsumfrage. Bertelsmann Stiftung (Hg.), Impuls Algorithmenethik #10, Gütersloh

Gučanin, J. (2018): Sexistischer AMS-Algorithmus benachteiligt Frauen und Mütter. In: Wienerin, 12.12.2018, wienerin.at/sexistischer-ams-algorithmus-benachteiligt-frauen-und-mutter (19.11.2020)

Hacker, P. (2018): Teaching Fairness to Artificial Intelligence: Existing and Novel Strategies Against Algorithmic Discrimination Under EU Law. In: Common Market Law Review 55(4), S. 1143–1186

Hagendorff, T. (2019a): Maschinelles Lernen und Diskriminierung: Probleme und Lösungsansätze. In: Österreichische Zeitschrift für Soziologie 44(S1), S. 53–66

Hagendorff, T. (2019b): Rassistische Maschinen? Übertragungsprozesse von Wertorientierungen zwischen Gesellschaft und Technik. In: Rath, M.; Krotz, F.; Karmasin, M. (Hg.): Maschinenethik. Normative Grenzen autonomer Systeme. Wiesbaden, S. 121–134

Hamilton, M. (2019): The sexist algorithm. In: Behavioral Sciences & the Law 37(2), S. 145–157

Hannah-Moffat, K.; Maurutto, P.; Turnbull, S. (2009): Negotiated Risk: Actuarial Illusions and Discretion in Probation. In: Canadian Journal of Law and Society 24(3), S. 391–409

Hayward, F. (2019): China unveils 500 megapixel camera that can identify every face in a crowd of tens of thousands. In: The Telegraph, 26.9.2019, www.telegraph.co.uk/news/2019/09/26/china-unveils-500-megapixel-camera-can-identify-every-face-crowd/ (19.11.2020)

Hoeren, T.; Niehoff, M. (2018): KI und Datenschutz – Begründungserfordernisse automatisierter Entscheidungen. In: RW Rechtswissenschaft 9(1), S. 47–66

Holl, J.; Kernbeiß, G.; Wagner-Pinter, M. (2018): Das AMS-Arbeitsmarktchancen-Modell. Dokumentation zur Methode. Synthesis Forschung GmbH (Hg.), Wien,

www.forschungsnetzwerk.at/downloadpub/arbeitsmarktchancen_methode_%20dokumentation.pdf (19.11.2020)
ITA (Institut für Technikfolgen-Abschätzung) (2019): Der AMS-Algorithmus am Prüfstand. (Allhutter, D.; Fischer, F.; Mager, A.), ITA-Dossier Nr. 43, Wien
Jaume-Palasí, L.; Spielkamp, M. (2017): Ethik und algorithmische Prozesse zur Entscheidungsfindung oder -vorbereitung. AlgorithmWatch Arbeitspapier Nr. 4, algorithmwatch.org/wp-content/uploads/2017/06/AlgorithmWatch_Arbeitspapier_4_Ethik_und_Algorithmen.pdf (19.11.2020)
Kessler, T.; Mummendey, A. (2007): Vorurteile und Beziehungen zwischen sozialen Gruppen. In: Jonas, K.; Stroebe, W.; Hewstone, M. (Hg.): Sozialpsychologie. Eine Einführung. Heidelberg, S. 487–531
Klare, B.; Burge, M.; Klontz, J.; Vorder Bruegge, R.; Jain, A. (2012): Face Recognition Performance: Role of Demographic Information. In: IEEE Transactions on Information Forensics and Security 7(6), S. 1789–1801
Kozyreva, A.; Herzog, S.; Lorenz-Spreen, P.; Hertwig, R.; Lewandowsky, S. (2020): Künstliche Intelligenz in Online-Umgebungen: Repräsentative Umfrage zur öffentlichen Meinung in Deutschland. Max-Planck-Institut für Bildungsforschung (Hg.), Berlin, https://pure.mpg.de/rest/items/item_3190264_9/component/file_3195146/content (19.11.2020)
Krafft, T.; Zweig, K. (2018): Wie Gesellschaft algorithmischen Entscheidungen auf den Zahn fühlen kann. In: Mohabbat Kar, R.; Thapa, B.; Parycek, P. (Hg.): (Un)berechenbar? Algorithmen und Automatisierung in Staat und Gesellschaft. Fraunhofer-Institut für Offene Kommunikationssysteme FOKUS, Berlin, S. 471–491
Krempl, S. (2019): »Algorithmen-TÜV«: Deutsches KI-Observatorium soll noch 2019 starten. In: heise online, 12.11.2019, heise.de/-4584781 (19.11.2020)
Lin, Z.; Jung, J.; Goel, S.; Skeem, J. (2020): The limits of human predictions of recidivism. In: Science Advances 6(7), doi: 10.1126/sciadv.aaz0652
Lobe, A. (2017): Der Algorithmus schlägt die letzte Stunde. In: Frankfurter Allgemeine Zeitung, 8.1.2017, www.faz.net/-gsf-8p2c5 (19.11.2020)
Mainzer, K. (2016): Künstliche Intelligenz – Wann übernehmen die Maschinen? Berlin/Heidelberg
Martini, M. (2018): Art. 22 DSGVO: Automatisierte Entscheidungen im Einzelfall einschließlich Profiling. In: Paal, B.; Pauly, D. (Hg.): Datenschutz-Grundverordnung, Bundesdatenschutzgesetz. München, S. 270–289
Martini, M. (2019): Blackbox Algorithmus – Grundfragen einer Regulierung Künstlicher Intelligenz. Berlin
Matzat, L.; Zielinski, L., Cocco, M., Penner, K., Spielkamp, M., Gießler, S., Lang, S., Thiel, V. (2019): Atlas der Automatisierung. Automatisierte Entscheidungen und Teilhabe in Deutschland. AlgorithmWatch (Hg.), Berlin, s100014241.ngcobalt360.manitu.net/atlas_algorithmwatch_org/wp-content/uploads/2019/07/Atlas_der_Automatisierung_von_AlgorithmWatch.pdf (19.11.2020)
Nordling, L. (2019): A fairer way forward for AI in health care. In: Nature 573(7775), S. S103–S105
Northpointe Inc. (2015): Practitioner's guide to COMPAS Core. https://assets.documentcloud.org/documents/2840784/Practitioner-s-Guide-to-COMPAS-Core.pdf (19.11.2020)
Obermeyer, Z.; Powers, B.; Vogeli, C.; Mullainathan, S. (2019): Dissecting racial bias in an algorithm used to manage the health of populations. In: Science 366(6464), S. 447–453

Orwat, C. (2020): Diskriminierungsrisiken durch Verwendung von Algorithmen. Baden-Baden
Orwat, C.; Schankin, A. (2018): Attitudes towards big data practices and the institutional framework of privacy and data protection – A population survey. KIT Scientific Reports 7753, Karlsruhe
Pohl, R. (2016): Cognitive Illusions: Intriguing phenomena in judgement, thinking and memory. Abingdon/New York
Puri, R. (2018): Mitigating Bias in AI Models. IBM, www.ibm.com/blogs/research/2018/02/mitigating-bias-ai-models/ (19.11.2020)
Raji, I.; Buolamwini, J. (2019): Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. In: Association for Computing Machinery: AIES '19: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. New York
Reichmann, W. (2019): Die Banalität des Algorithmus. In: Rath, M.; Krotz, F.; Karmasin, M. (Hg.): Maschinenethik. Normative Grenzen autonomer Systeme. Wiesbaden, S. 135–153
Rich, A.; Gureckis, T. (2019): Lessons for artificial intelligence from the study of natural stupidity. In: Nature Machine Intelligence 1(4), S. 174–180
Sandvig, C.; Hamilton, K.; Karahalios, K.; Langbort, C. (2016): When the Algorithm Itself Is a Racist: Diagnosing Ethical Harm in the Basic Components of Software. In: International Journal of Communication 10, S. 4972–4990
Saurwein, F. (2019): Automatisierung, Algorithmen, Accountability. Eine Governance Perspektive. In: Rath, M.; Krotz, F.; Karmasin, M. (Hg.): Maschinenethik. Normative Grenzen autonomer Systeme. Wiesbaden, S. 35–56
Schaar, P. (2017): Überwachung, Algorithmen und Selbstbestimmung. In: Gapski, H.; Oberle, M.; Staufer, W. (Hg.): Medienkompetenz. Herausforderung für Politik, politische Bildung und Medienbildung. Bundeszentrale für politische Bildung, Schriftenreihe Band 10111, Bonn, S. 73–81
Schauer, F. (2018): Statistical (And Non-)Statistical Discrimination. In: Lippert-Rasmussen, K. (Hg.): The Routledge Handbook of the Ethics of Discrimination. London, S. 42–53
Scherr, A. (2016): Diskriminierung. Wie Unterschiede und Benachteiligungen gesellschaftlich hergestellt werden. Wiesbaden
Scholz, P. (2019): Artikel 22 DSGVO: Automatisierte Entscheidungen im Einzelfall einschließlich Profiling. In: Simitis, S.; Hornung, G.; Spiecker genannt Döhmann, I. (Hg.): Datenschutzrecht. DSGVO mit BDSG. Baden-Baden, S. 704–721
Schwär, H. (2019): Jobcenter setzen auf Künstliche Intelligenz – die Folgen für Bewerber könnten fatal sein. In: Business Insider, 3.4.2019, www.businessinsider.de/tech/jobcenter-setzen-laut-forschern-auf-kuenstliche-intelligenz-die-folgen-fuer-bewerber-koennten-fatal-sein-2019-4/ (19.11.2020)
Seeger, C. (2017): Einsatz und Einfluss von Algorithmen auf das digitale Leben. Wissenschaftliche Dienste des Deutschen Bundestages, Aktueller Begriff Nr. 26/17, Berlin
Snow, J. (2018): Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots. American Civil Liberties Union (Hg.), New York, www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28 (19.11.2020)
Streim, A.; Dehmel, S. (2018): Künstliche Intelligenz: Bundesbürger sehen vor allem Chancen. Presseinformation vom 27.11.2018, Bitkom, www.bitkom.org/Presse/

Presseinformation/Kuenstliche-Intelligenz-Bundesbuerger-sehen-vor-allem-Chancen (19.11.2020)
Sutmöller, N. (2019): Big Data und die Frage nach Gerechtigkeit. In: Rath, M.; Krotz, F.; Karmasin, M. (Hg.): Maschinenethik. Normative Grenzen autonomer Systeme. Wiesbaden, S. 155–172
Szigetvari, A. (2018): AMS-Vorstand Kopf: »Was die EDV gar nicht abbilden kann, ist die Motivation«. In: Der Standard, 10.10.2018, www.derstandard.at/story/2000089096795/ams-vorstand-kopf-menschliche-komponente-wird-entscheidend-bleiben (19.11.2020)
Vanwesenbeeck, I. (2017): Sex Work Criminalization Is Barking Up the Wrong Tree. In: Archives of Sexual Behavior 46(6), S. 1631–1640
Vieth, K.; Wagner, B. (2017): Teilhabe, ausgerechnet. Wie algorithmische Prozesse Teilhabechancen beeinflussen können. Bertelsmann Stiftung (Hg.), Impuls Algorithmenethik #2, Gütersloh
Vogd, W. (2004): Ärztliche Entscheidungsfindung im Krankenhaus. Komplexe Fallproblematiken im Spannungsfeld von Patienteninteressen und administrativ-organisatorischen Bedingungen. In: Zeitschrift für Soziologie 33(1), S. 26–47
Volanen, S.-M.; Suominen, S.; Lahelma, E.; Koskenvuo, M.; Silventoinen, K. (2006): Sense of coherence and its determinants: a comparative study of the Finnish-speaking majority and the Swedish-speaking minority in Finland. In: Scandinavian journal of public health 34(5), S. 515–525
Wagner, G. (2019): Künstliche Intelligenz verhindert Diskriminierung? Muss nicht – kann aber! In: DIW Wochenbericht 86(20), S. 372
Weichert, T. (2018): Big Data im Gesundheitsbereich. ABIDA – Assessing Big Data. www.abida.de/sites/default/files/ABIDA%20Gutachten-Gesundheitsbereich.pdf (19.11.2020)
Weitzer, R. (2018): Resistance to sex work stigma. In: Sexualities 21(5-6), S. 717–729
Wersig, M. (2017): Fälle zum Allgemeinen Gleichbehandlungsgesetz (AGG). Eine Einführung in Theorie und Praxis des Antidiskriminierungsrechts in 22 Fällen. Opladen/Toronto
Wilson, B.; Hoffman, J.; Morgenstern, J. (2019): Predictive Inequity in Object Detection. Cornell University, https://arxiv.org/pdf/1902.11097.pdf (19.11.2020)
Wimmer, B. (2018a): AMS-Chef: »Mitarbeiter schätzen Jobchancen pessimistischer ein als der Algorithmus«. In: futurezone, 12.10.2018, futurezone.at/netzpolitik/ams-chef-mitarbeiter-schaetzen-jobchancen-pessimistischer-ein-als-der-algorithmus/400143839 (19.11.2020)
Wimmer, B. (2018b): Der AMS-Algorithmus ist ein »Paradebeispiel für Diskriminierung«. In: futurezone, 17.10.2018, futurezone.at/netzpolitik/der-ams-algorithmus-ist-ein-paradebeispiel-fuer-diskriminierung/400147421 (19.11.2020)
Wimmer, B. (2018c): »AMS-Sachbearbeiter erkennen nicht, wann ein Programm falsch liegt«. In: futurezone, 18.10.2018, futurezone.at/netzpolitik/ams-sachbearbeiter-erkennen-nicht-wann-ein-programm-falsch-liegt/400147472 (19.11.2020)
YVTltk (Yhdenvertaisuus- ja tasa-arvolautakunta) (2018): Assessment of creditworthiness, authority, direct multiple discrimination, gender, language, age, place of residence, financial reasons, conditional fine. National Non-Discrimination and Equality Tribunal of Finland/Plenary session (voting). Register number: 216/2017, www.yvtltk.fi/en/index/opinionsanddecisions/decisions.html (19.11.2020)
Zgoll, M. (2019): Wegen seiner Adresse: Darum bekommt Herr Wenskowski keinen Mobilfunkvertrag. In: Hannoversche Allgemeine Zeitung, 12.9.2019,

https://www.haz.de/Hannover/Aus-der-Stadt/Nedderfeldstrasse-in-Linden-Nord-Wie-Herr-Wenskowski-wegen-seiner-Adresse-keinen-Mobilfunkvertrag-bekommt (19.11.2020)

Zweig, K.; Krafft, T. (2018): Fairness und Qualität algorithmischer Entscheidungen. In: Mohabbat Kar, R.; Thapa, B.; Parycek, P. (Hg.): (Un)berechenbar? Algorithmen und Automatisierung in Staat und Gesellschaft. Berlin, S. 204–227

Zweig, K.; Wenzelburger, G.; Krafft, T. (2018): On Chances and Risks of Security Related Algorithmic Decision Making Systems. In: European Journal for Security Research 3(2), S. 181–203

BÜRO FÜR TECHNIKFOLGEN-ABSCHÄTZUNG BEIM DEUTSCHEN BUNDESTAG

Karlsruher Institut für Technologie

Neue Schönhauser Straße 10
10178 Berlin

Telefon: +49 30 28491-0
E-Mail: buero@tab-beim-bundestag.de
Web: www.tab-beim-bundestag.de
Twitter: @TABundestag