Antidiskriminierungsstelle
des Bundes

Diskriminierungsrisiken durch Verwendung von Algorithmen

Carsten Orwat

Nomos

Diskriminierungs-risiken durch Verwendung von Algorithmen

Eine Studie, erstellt mit einer Zuwendung der Antidiskriminierungsstelle des Bundes.

von Dr. Carsten Orwat

Institut für Technikfolgenabschätzung und Systemanalyse (ITAS)

Karlsruher Institut für Technologie (KIT)

Inhaltsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

Abs.	Absatz
ACLU	American Civil Liberties Union
ADS	Antidiskriminierungsstelle des Bundes
AGG	Allgemeines Gleichbehandlungsgesetz
AI	Artificial intelligence
AMS	Arbeitsmarktservice
Art.	Artikel
BGH	Bundesgerichtshof
BDSG	Bundesdatenschutzgesetz
BDSG a. F.	Bundesdatenschutzgesetz, alte Fassung
BDSG n.F.	Bundesdatenschutzgesetz, neue Fassung
BVerfG	Bundesverfassungsgericht
BVerfGE	Entscheidung des Bundesverfassungsgerichts
bzw.	beziehungsweise

ca.	circa
CNIL	Commission Nationale de l'Informatique et des Libertés
DDD	Défenseur des Droits
d. h.	das heißt
DM	Data-Mining
DSGVO	Datenschutz-Grundverordnung
DSRL	Datenschutzrichtlinie
ebd.	ebenda
EDPB	European Data Protection Board
EMRK	Europäische Menschenrechtskonvention
Erwg.	Erwägungsgrund
etc.	et cetera
EU	Europäische Union
GG	Grundgesetz
GRCh	EU-Grundrechte-Charta
HUD	U.S. Department of Housing and Urban Development
IT	Informationstechnologie
KI	Künstliche Intelligenz
ML	Maschinelles Lernen bzw. „machine learning“
NFHA	National Fair Housing Alliance
PoC	People of Color

ROC AUC	Receiver Operating Characteristic, Area under the Curve
RL	Richtlinie
Rn.	Randnummer
s. o.	siehe oben
s. u.	siehe unten
u. a.	und andere oder unter anderem
WP29	Article 29 Data Protection Working Party, umbenannt in European Data Protection Board (EDPB)
YVTltk	Yhdenvertaisuus- ja tasa-arvolautakunta
z. B.	zum Beispiel

Danksagung und Finanzierung

Die Erstellung der Studie wurde mit einer Zuwendung der Antidiskriminierungsstelle des Bundes (ADS) gefördert, auf die an dieser Stelle mit großem Dank hingewiesen werden soll.

Der Autor wird ansonsten aus Mitteln der institutionellen Förderung der Helmholtz-Gemeinschaft über das Institut für Technikfolgenabschätzung und Systemanalyse finanziert. Die Helmholtz-Gemeinschaft wird durch Bund und Länder finanziert.

Frau Nathalie Schlenzka von der Antidiskriminierungsstelle des Bundes (ADS) stellte mit einer Abfrage an europäische Antidiskriminierungsstellen hilfreiche Informationen und Kontakte zur Verfügung. Ihr und ihren Kolleg*innen danke ich für fruchtbare Kommentare. An dieser Stelle sei auch den Mitarbeiter*innen der europäischen Antidiskriminierungsstellen gedankt, die Fragen zu Diskriminierungsfällen sowie zu ihrer Arbeitsweise und ihren Kompetenzen beantwortet haben. Die Antworten sind an den zitierten Stellen in die Studie aufgenommen worden.

Für zahlreiche intensive und äußerst fruchtbare Gespräche danke ich Oliver Raabe, Moritz Renftle, Oliver Siemoneit und Reinhard Heil. Oliver Raabe hat zudem dankenswerter Weise die Durchsicht auf rechtswissenschaftliche Fragestellungen geleistet. Arnold Sauter, Christoph Kehl und Christian Wadephul möchte ich für hilfreiche Kommentare danken. Eventuell verbleibende Fehler sind selbstverständlich die des Autors.

Hinweis

Die hier genannten Einschätzungen, Äußerungen und Meinungen sind allein die des Autors und geben nicht notwendigerweise die offizielle Meinung der Antidiskriminierungsstelle des Bundes wieder.

Zusammenfassung

Algorithmen: Im Fokus der Studie stehen Algorithmen, die zur Datenverarbeitung und halb- oder vollautomatisierten Ausführung von Entscheidungsregeln für die Differenzierung von Personen eingesetzt werden. Differenzierungen beziehen sich dabei auf wirtschaftliche Produkte, Dienste, Positionen oder Entgelte sowie auf staatliche Entscheidungen und Handlungen, die individuelle Freiheiten oder die Verteilung von Leistungen betreffen.

Diskriminierung: Aus algorithmenbasierten Differenzierungen werden insbesondere dann Diskriminierungen, wenn sie eine ungerechtfertigte Benachteiligung von Personen darstellen, die durch geschützte Merkmale (insbesondere Alter, Geschlecht, ethnische Herkunft, Religion, sexuelle Orientierung oder Behinderung) gekennzeichnet sind. In der Studie werden Beispielfälle beschrieben, in denen algorithmen- und datenbasierte Differenzierungen als Diskriminierung rechtlich festgestellt wurden oder die als Diskriminierungsrisiken analysiert und diskutiert werden.

Ersatzinformationen: Algorithmen- und datenbasierte Differenzierungen weisen häufig Charakteristika des Typs der sogenannten statistischen Diskriminierung auf. Dabei werden zur Differenzierung Ersatzinformationen bzw. -variablen oder Proxies (z. B. Alter) herangezogen, weil die eigentlichen Eigenschaften, nach denen differenziert werden soll (z. B. Arbeitsproduktivität), für die Entscheidenden durch Einzelfallprüfungen nur schwer ermittelbar sind. Diese Ersatzvariablen können geschützte Merkmale sein oder Korrelationen zu ihnen aufweisen. Mit algorithmischen Verfahren des Data-Minings und des maschinellen Lernens können anstelle einer oder weniger Ersatzvariablen auch komplexe Modelle mit einer Vielzahl von Variablen treten.

Gesellschaftliche Risiken: Die Legitimität derartiger Differenzierungen wird mit Effizienzvorteilen bei der Überwindung von Informationsdefiziten begründet, doch gleichzeitig bergen sie gesellschaftliche Risiken durch mögliches Generalisierungsunrecht, Behandlung von Menschen als bloße Mittel,

Einschränkung der freien Entfaltung der Persönlichkeit, Akkumulations- und Verstärkungseffekte von gesellschaftlichen Ungleichheiten oder Risiken für Gleichheits- oder sozialpolitische Ziele. Viele Diskriminierungsrisiken bei der Entwicklung und Verwendung von Algorithmen resultieren aus der Verwendung von Daten, die frühere Ungleichbehandlungen abbilden.

Bedarf nach gesellschaftlichen Abwägungen: Die Lösung von technisch bedingten Diskriminierungsrisiken von Algorithmen ist zwar grundlegend, doch die Legitimität von Anwendungen verschiedener Formen der algorithmenbasierten Differenzierungen an sich bedarf gesellschaftlicher Abwägungen und Festlegungen, die an den genannten gesellschaftlichen Risiken ansetzen und die Differenzierungs- und Effizienzgewinne, aber vor allem deren gesellschaftliche Verteilung berücksichtigen und zu Festlegungen über für die Gesellschaft akzeptable Differenzierungsanwendungen kommen sollten. Da in den meisten Fällen derartige Differenzierungen auf Basis der Auswertung umfassender Mengen an personenbezogenen Daten erfolgen, sind ebenso Risiken für das Recht auf informationelle Selbstbestimmung einzubeziehen.

Datenschutzrecht: Das gegenwärtige Datenschutzrecht bedarf Klarstellungen und Korrekturen, die auch der Antidiskriminierung dienen würden. Diese beziehen sich unter anderem auf die sogenannte informierte Einwilligung, bei der die Betroffenen die weitreichenden Konsequenzen, auch in Form von möglichen Ungleichbehandlungen, zum Zeitpunkt der Einwilligung abschätzen müssen. Das scheint jedoch angesichts tatsächlicher Praktiken der Erfassung, Weitergabe, Verknüpfung und Nutzung von personenbezogenen Daten nicht mehr adäquat.

Detaillierung der Regulierung von Entscheidungen: Des Weiteren wird eine Ergänzung des Regulierungsfokus nicht nur auf der Datenverarbeitung, sondern auch auf einer detaillierteren Regulierungen der algorithmen- und datenbasierten Entscheidungen vorgeschlagen. Das Antidiskriminierungsrecht mit Vorgaben zu den geschützten Merkmalen, die bei bestimmten Entscheidungssituationen verwendet werden dürfen, stellt bereits eine Regulierung von Entscheidungen dar. Verbesserungen des Regulierungsrahmens können von einer detaillierten Spezifizierung der erlaubten Verwendung bestimmter Entscheidungsmerkmale, z. B. über Konkretisierun-

gen bei Ausnahmen nach sachlichem Grund und sogenannten anerkannten Berechnungsmethoden bis hin zum Verbot bestimmter algorithmen- und datenbasierter Differenzierungen für bestimmte Entscheidungen mit hohem Risiko reichen. Das datenschutzrechtliche Verbot von automatisierten Entscheidungen kann an verschiedenen Stellen verbessert werden. Regulatorische Instrumente sollten nach dem Ausmaß der gesellschaftlichen Risiken der Differenzierungen unterschiedlich gestaltet werden.

Aufgaben und Pflichten der Antidiskriminierungsstellen: Insbesondere die schwierige Nachweisbarkeit von algorithmen- und datenbasierten Diskriminierungen durch die Betroffenen selbst legen ein Vorgehen nach dem Subsidiaritätsprinzip durch stellvertretende Einrichtungen und dem kollektiven Rechtsschutz nahe.

- Viele Beispiele verdeutlichen, dass Identifizierung und Nachweis von algorithmenbasierten Diskriminierungen auch ohne die direkte Inspektion des Algorithmus bzw. der „Öffnung" des Softwaresystems erfolgen können. Stattdessen erfolgt dort der Nachweis über die Erfassung und Auswertung öffentlich zugänglicher Daten zu den Ergebnissen der Differenzierungsentscheidungen, die zu den Interaktionen und Transaktionen der untersuchten Dienste und Produkte ermittelbar waren. Wo die Ergebnisse nicht ermittelbar sind, stößt ein solches Vorgehen jedoch an Grenzen.

- An Untersuchungen der Ergebnisse algorithmenbasierter Entscheidungen und an Anfragen und Hinweisen von Betroffenen oder aus den Medien können auch die Aufgaben der Antidiskriminierungsstellen, d. h. die Beratung und Unterstützung von Betroffenen von Benachteiligungen ansetzen, auch im Onlinebereich oder bei computergestützten Entscheidungen. Sind allerdings Informationen zu den Ergebnissen von Entscheidungen nicht öffentlich verfügbar, dann sind weitergehende Zugangsmöglichkeiten und -rechte zu relevanten Informationen für die Antidiskriminierungsstellen erforderlich, wenn sie entsprechend ihrem Auftrag Diskriminierungen identifizieren und vermindern sollen.

- Bei durch Algorithmen ermöglichten, personalisierten Angeboten und Diensten kann es für die Betroffenen schwierig sein, Differenzierungen von Personen überhaupt zu erkennen und notwendige Vergleiche für den Nachweis von ungerechtfertigten Benachteiligungen zu erbringen. Antidiskriminierungsstellen haben üblicherweise

dazu Kenntnisse und Erfahrungen über diskriminierungsanfällige Personengruppen, Situationen oder Behandlungsformen, den Diskriminierungsursachen, scheinbar neutralen Kriterien und über Korrelationen zu geschützten Merkmalen. Diese Kenntnisse können auch die Ausgangspunkte für systematische empirische Untersuchungen und Testings sein, die zum Auftrag der Antidiskriminierungsstellen gehören.

- Insbesondere bei Algorithmen der künstlichen Intelligenz und bei Anwendungen von automatisierten Entscheidungen kann es durch rechtliche Vorgaben für die Anwendenden erforderlich werden, dass sie potenzielle Diskriminierungsrisiken abschätzen, die Erklärbarkeit der Algorithmen sicherstellen sowie die Funktionsweisen von Algorithmen, der Entscheidungsregeln und deren Auswirkungen auf Betroffene – auch in Form von möglichen Diskriminierungen – dokumentieren müssen. Zu diesen Dokumentationen sollten Antidiskriminierungsstellen in Fällen des Verdachts auf Diskriminierung Zugang erhalten, wobei das Zugangsrecht gesetzlich zu regeln ist.

- Weitere (potenzielle) Aufgaben von Antidiskriminierungsstellen umfassen die Beratung von Entwickelnden und Anwendenden zur präventiven Vermeidung von Diskriminierungen und die (verpflichtende) Einbindung in öffentliche Beschaffungsvorgänge von algorithmenbasierten Systemen, die besonders anfällig für Diskriminierungsrisiken sind.

Summary

Algorithms: The study focuses on algorithms that are used for data processing and semi- or full-automated implementations of decision rules to differentiate between individuals. Such differentiations relate to commercial products, services, positions or payments as well as to state decisions and actions that affect individual freedoms or the distribution of services.

Discriminations: Algorithm-based differentiations become discriminatory if they lead to unjustified disadvantaging of persons with legally protected characteristics, in particular age, sex, ethnic origin, religion, sexual orientation, or disability. The study describes cases in which algorithm- and data-based differentiations have been legally classified as discrimination or which are analysed or discussed as risks of discrimination.

Surrogate information: Algorithm- and data-based differentiations often exhibit the characteristics of so-called statistical discrimination. Typical for this kind of discrimination is the use of surrogate information, surrogate variables or proxies (e.g. age) to differentiate, because the original distinguishing characteristics (e.g. labour productivity) are difficult for the decision-makers to determine by examining individual cases. These surrogate variables can be protected characteristics, or there can be correlations between them and protected characteristics. With algorithmic methods of data mining and machine learning, complex models with a large number of variables can be used instead of one or a few surrogate variables.

Societal risks: The legitimacy of such differentiations is often justified on the grounds of efficiency in overcoming information deficits. However, they also involve societal risks such as injustice by generalisation, treatment of humans as mere objects, restriction of the free development of personality, accumulation effects and growing inequality and risks to societal goals of equality and social policy. Many discrimination risks of developing and using algorithms result from the use of data reflecting former unequal treatments.

Needs for societal considerations: Although overcoming technically-based discrimination risks of algorithms and data is fundamental, the legitimacy of different forms of algorithmic differentiation requires societal considerations and decisions that take into account the aforementioned societal risks, the benefits of differentiation and, in particular, their distribution in society. This should lead to definitions of socially acceptable differentiations. Since in most cases such differentiations are based on the processing of comprehensive amounts of personal data, risks to the right to informational self-determination must be considered as well.

Data protection law: The current data protection law needs clarifications and corrections that would also serve anti-discrimination purposes. These relate, among other things, to the so-called informed consent, where affected persons must assess far-reaching potential consequences, including possible unequal treatments, at the time of consent. This approach no longer seems adequate in light of the actual practices of collecting, merging, transferring, and using personal data and the emerging risks of unequal treatments based on these practices.

More detailed regulation of decisions: Furthermore, it is suggested that the current regulatory focus on data processing should be supplemented by a focus on the decision-making based on algorithms. Anti-discrimination law, with its provisions on protected characteristics that may, or shall not, be used in certain decision-making situations, can already been seen as a regulation of decisions. Improvements in regulation can range from more detailed provisions on the permitted use of certain decision criteria, e.g. by clarifying exemptions from the prohibited use justified by legitimate aims or by the use of recognised methods, to the prohibition of certain algorithm-based differentiations for certain types of decisions with high risks. Regulatory instruments should be designed according to the specific level of societal risks attributable to different types of algorithm-based differentiations.

Tasks and duties of equality bodies: In particular, the difficulty in detecting and proving algorithm-based discrimination by affected persons suggests, according to the principle of subsidiarity, that representative bodies should

take action on behalf of the affected persons and collective redress should be used.

- Many examples illustrate that detecting and proving discrimination with algorithms is also possible without direct inspection of the algorithm or "opening" the software system. Instead, evidence of unequal treatment or discrimination can be provided by collecting and investigating publicly available data on the outcomes of differentiation decisions, which are derived from the interactions and transactions of the services and products investigated.

- Together with requests and information from those affected and the media, such investigations of the outcomes can serve as starting points for the equality bodies' tasks of advising and supporting persons affected by discrimination, also for the online area and automated decision-making. If there is not enough information publicly available on the outcomes of decisions, the access options and rights of equality bodies should be extended so that they can fulfil their mandate to identify and reduce discrimination.

- For personalized offers and services, enabled by algorithms, it can be difficult for affected persons to detect differentiations of persons and to make comparisons in order to provide evidence for unequal treatments. Equality bodies generally have expertise and experiences with regard to groups of persons, situations and treatments prone to discrimination, the usual rationales for discrimination, seemingly neutral criteria and correlations with protected characteristics. Such expertise can be starting points for systematic empirical investigations, anti-discrimination testings and algorithm audits.

- In particular, the use of artificial intelligence algorithms and applications in automated decision-making can require by legal provision that entities using them assess discrimination risks, document, among other things, their functioning and decision rules, and ensure explainability also with regard to possible consequences including unequal treatment. Such documentation should be accessible to equality bodies in cases of suspected discrimination, with the right of access being regulated by law.

- Other (potential) tasks of equality bodies include advising entities developing and implementing algorithms on the prevention of discrimination and (mandatory) involvement in public procurement procedures of algorithm-based systems that are particularly prone to discrimination.

1. Einleitung

Algorithmen und umfangreiche Datensätze sind zunehmend bei Entscheidungen involviert, die nicht nur triviale Konsequenzen für Menschen haben, sondern vielmehr ihre Lebensführung und Persönlichkeitsentfaltung beeinflussen. Algorithmen produzieren dabei Schlussfolgerungen und Ergebnisse, die menschliche Entscheidende als Informationsgrundlage ihrer Entscheidungen nutzen, oder aber die Ausführung von Entscheidungsregeln wird an Algorithmen bzw. die sie beinhaltenden Computersysteme vollständig delegiert.

Abgesehen von einigen frühen Analysen zu Bias und Ungleichbehandlungen durch Verwendung von Computersystemen (Friedman & Nissenbaum 1996; Bruce & Adam 1989) sind Diskriminierungsrisiken im Zusammenhang mit Informations- und Kommunikationstechnologien erst in diesem Jahrzehnt umfassend thematisiert worden. Vor allem im Zuge der Big-Data-Entwicklung wurde auf damit verbundene Diskriminierungsrisiken hingewiesen (z. B. Crawford 2013; Dwork & Mulligan 2013; The White House 2014; US CEA 2015; FTC 2016; Schneider & Ulbricht 2018). Zahlreiche Beispielfälle zu Ungleichbehandlungen und Diskriminierungen durch die Anwendung von Algorithmen wurden durch Forschungen oder Journalist*innen aufgedeckt, auf die im Folgenden eingegangen werden wird.

Die vorliegende Studie hat eine problemorientierte Sicht auf die Folgen der Verwendung von Algorithmen bei der Differenzierung von Personen. Im Fokus der Studie stehen Algorithmen, die für die Differenzierung von Personen im Hinblick auf ebenso differenzierte Informationen, Produkte, Dienste, Entgelte, Positionen etc. eingesetzt werden.[1] Derartige Differenzierungen können zu ungerechtfertigten Ungleichbehandlungen bzw. Diskriminierungen führen. Dabei ist es eine gesellschaftliche Aufgabe, zu

[1] Durch die problemorientierte Sichtweise bleiben die zahlreichen diskriminierungsfreien Anwendungen von Algorithmen unberücksichtigt. Die Studie behandelt auch nicht die Auswirkungen von Systemen, die über die algorithmenbasierte Steuerung von Informationen Auswirkungen auf die Informationswahrnehmung („Filterblasen"), Meinungsfreiheit, Meinungsbildung oder das Wahlverhalten in demokratischen Prozessen haben. Ebenso werden zwar Beispiele aus dem Bereich staatlichen Handelns genannt, aber die Schlussfolgerungen, Ableitungen von Handlungsbedarfen oder -optionen werden vor allem mit Blick auf den privatwirtschaftlichen Bereich gezogen.

definieren, welche Ungleichbehandlung als ungerechtfertigt gilt und ungerechtfertigte Ungleichbehandlungen zu regulieren. Diese Aufgabe betrifft auch die Differenzierungsformen, die mit der Verwendung von Algorithmen realisiert werden.

Nach einer kurzen Einführung zu Algorithmen, relevanten Entwicklungen in der Datenverarbeitung und algorithmenbasierten Differenzierungen (Kapitel 2), werden relevante Typen von Diskriminierung dargestellt (Kapitel 3). Dabei spiegeln Diskriminierungen, die durch algorithmen- und datenbasierte Differenzierungen verursacht werden, vor allem den Typ der statistischen Diskriminierung wider, was auch an den geschilderten Beispielen deutlich wird (Kapitel 4). Die Ursachen für Diskriminierungsrisiken sind dabei nicht nur bei der Art und Weise der Erzeugung, Auswahl und Verwendung von Algorithmen und Datensätzen zu sehen, sondern auch in der Verwendung von Differenzierungen an sich (Kapitel 5). Anschließend werden Überlegungen zu Handlungsbedarfen und -optionen angestellt, die vor allem der Diskriminierungsvermeidung dienen sollen, und darüber hinausgehende gesellschaftliche Abwägungsprozesse gefordert (Kapitel 6).

2. Begriffe und grundlegende Entwicklungen

2.1 Algorithmen

In dieser Studie wird der Begriff „Algorithmus“ aus informationstechnischer Perspektive gebraucht.[2] Danach sind mit Algorithmen grundlegende, formalisierte und präzise festgelegte Berechnungsvorschriften bzw. Regeln für eine Abfolge von Berechnungsschritten gemeint, die eine vorgegebene Aufgabe bewältigen sollen. Für eine berechenbare Aufgabe, wie beispielsweise das Sortieren von Listen, gibt es oft sehr viele verschiedene Algorithmen.

Algorithmen müssen zu ihrer Ausführung durch einen Computer in eine der zahlreichen Programmiersprachen (Python, Java, JavaScript, C++ etc.) implementiert bzw. programmiert werden, danach liegen sie als Programmteile vor, die zusammen mit Datenstrukturen zu Software bzw. Softwaresystemen kombiniert werden. Dort erfüllen Algorithmen die Aufgabe, aus einem Input, meist in Form von Daten, einen Output, meist in anderen Datenformaten, zu erzeugen. Wird im Folgenden von Algorithmen gesprochen, so sind immer Implementierungen von Algorithmen als mögliche Bestandteile von Software gemeint.

Algorithmen werden in Software implementiert, kombiniert und organisiert, um bestimmte Zwecke zu erfüllen. Diese Zwecksetzungen werden vor allem durch Menschen, insbesondere Softwareentwickelnde und Auftraggebende, vorgegeben, ebenso die in Algorithmen beinhalteten Interpretationen, Wertsetzungen, Priorisierungen und Ausschlüsse.[3] Bei ihren Zwecksetzungen und Anwendungen können auch unintendierte Folgen[4] sowohl für die direkt Betroffenen als auch für indirekt betroffene Dritte auftreten, die die Realisierung von gesellschaftlichen und grundrechtlich geschützten Werten wie den Schutz der Menschenwürde, die Wahrung der freien

[2] Zur Diskussion des Begriffs Algorithmus und dessen Abgrenzung siehe z. B. Hill (2016), Cormen u. a. (2010: 5-15), Mittelstadt u. a. (2016) oder Yeung (2017).
[3] Vgl. z. B. Schinzel (2017), Zweig, Fischer & Lischka (2018) oder Kitchin (2017).
[4] Hier nach Brey (2000), (2009) und Kitchin (2017).

Entfaltung der Persönlichkeit und der informationellen Selbstbestimmung, die Vermeidung von Diskriminierungen oder die Sicherung der Rechtsstaatlichkeit erschweren. Algorithmen entfalten erst durch ihren Einsatz in Softwareanwendungen unter Verwendung bestimmter Datensätze in wirtschaftlichen, sozialen, administrativen und rechtlichen Praktiken ihre gesellschaftlichen Konsequenzen. Daher richtet sich in der Studie der Blick weniger auf Algorithmen an sich, als vielmehr auf ihre Anwendungen in Softwaresystemen für bestimmte Zwecksetzungen.

Dass derzeit Algorithmen, vor allem in den Gesellschafts- und Geisteswissenschaften, der Öffentlichkeit und der Politik, viel Aufmerksamkeit gewidmet wird, kann dadurch erklärt werden, dass (a) IT-Systeme mit Algorithmen nicht nur in Produktionsprozessen, Büroanwendungen und wirtschaftlichen Transaktionen, sondern auch in sozialen Interaktionen und Kommunikation nahezu ubiquitär in allen Lebensbereichen eingesetzt werden, (b) ihre verbreitete Nutzung für die automatisierte Bewältigung von großen Datenmassen, beschleunigten Interaktionen und Transaktionen unerlässlich und teilweise durch wirtschaftliche Netzwerkeffekte[5] unausweichlich erscheint, (c) sie vermehrt bei Entscheidungen, die folgenreich die Lebenschancen und Persönlichkeitsentfaltung von Menschen betreffen, eingesetzt werden und (d) ihnen teilweise eine gewisse Handlungs- und Entscheidungsfähigkeit zugemutet oder sogar eine Verlagerung von Verantwortung auf Algorithmen angedeutet wird.

Für die vorliegende Studie sollen Algorithmen in mehrere **Typen** unterschieden werden, zunächst in[6] (1) Algorithmen, deren Regeln gänzlich durch menschliche Logik entwickelt werden und deren Regeln als „direkte Programmierung“ quasi „per Hand“ von den Entwickelnden umgesetzt

---

[5] Bei ökonomischen Netzwerkeffekten steigt der Nutzen für eine*n einzelne*n Nutzer*in mit der Anzahl von weiteren Nutzenden, weil sich insbesondere die Kommunikations- oder Austauschmöglichkeiten zwischen den Nutzenden erhöht. Dadurch kann es zu Konzentrations- oder Monopolisierungstendenzen bzw. zur Dominanz eines Systems kommen. Für die oder den einzelne*n Nutzer*in können sich, wenn sie oder er an der Kommunikation oder an dem Austausch teilnehmen will, die Wahlmöglichkeiten drastisch reduzieren und die Nutzung des einen Systems kann quasi unausweichlich werden. Netzwerkeffekte treten insbesondere bei Telekommunikationsnetzwerken, bei Plattformen des „match-making“ (z. B. Plattformen der Arbeitsvermittlung, Handelsplattformen wie eBay, Dating-Plattformen) oder des „audience-making“ (z. B. soziale Netzwerke, Suchmaschinen), bei Transaktionssystemen (z. B. elektronische Bezahlsystemen wie PayPal) und bei Software-Plattformen (insbesondere Betriebssysteme) auf. Für eine Übersicht und Diskussion siehe z. B. Dewenter & Lüth (2018).

[6] Hier nach Lehr & Ohm (2017), Selbst & Barocas (2018).

werden, und (2) Algorithmen des Data-Minings bzw. maschinellen Lernens, deren Regeln auf Korrelationen beruhen, die durch die Auswertung von Daten erzeugt werden. Letztere werden auch als „lernende Algorithmen" bezeichnet und üblicherweise dem Bereich der künstlichen Intelligenz zugeordnet (Näheres in Abschnitt 2.2.2 und 3.4).[7] Auch bei diesem Typ findet das „Lernen" nicht ohne Menschen statt, da Entwickelnde und Anwendende bei Verfahren des maschinellen Lernens viele Gestaltungsentscheidungen treffen müssen. Dabei sind in vielen Fällen die menschlichen Entscheidungen auch die Quellen von Diskriminierungsrisiken (Näheres in Kapitel 4 und 5).

Algorithmen werden heute nicht nur zur Automatisierung der Datenverarbeitung einschließlich Datenanalyse und Ableitung von Schlussfolgerungen eingesetzt, sondern auch um Entscheidungsregeln automatisiert anzuwenden und durchzusetzen. Daher können zur Veranschaulichung Algorithmen in einer weiteren Unterteilung in (1) solche der automatischen Datenverarbeitung und -analyse und in (2) solche, die Entscheidungsregeln automatisiert vollziehen[8], unterschieden werden (Ernst 2017; Kleinberg u. a. 2019). Viele der hier interessierenden Softwaresysteme beinhalten beides, für die späteren Betrachtungen ist es aber sinnvoll, sie in ihrer Darstellung zu unterscheiden. Denn nicht jede Anwendung von Algorithmen bedeutet eine vollautomatisierte Entscheidung. Häufig werden Algorithmen auch lediglich zur Datenanalyse genutzt, deren Ergebnisse als Empfehlungen oder Unterstützung von menschlichen Entscheidungen verwendet werden.

---

[7] Leider hat sich in der Diskussion über Algorithmen ergeben, dass oft nur noch die Algorithmen des maschinellen Lernens gemeint sind, wenn von „Algorithmen" die Rede ist. Dem wird aber in dieser Studie nicht gefolgt, da sich Diskriminierungsrisiken auch durch die Verwendung einer Vielzahl von Algorithmen ergeben können.

[8] Näheres in Abschnitt 2.3.3.

2.2 Entwicklungen bei der Datenverarbeitung

2.2.1 Ausweitung der Menge an personenbeziehbaren Daten

In den letzten Jahrzehnten ist die Menge an personenbeziehbaren und personenbezogenen Daten, die als Produkt oder Nebenprodukt der Computerisierung (heute vermehrt auch „Digitalisierung" genannt) nicht nur von und zwischen Organisationen, sondern auch in öffentlichen und privaten Lebensbereichen erzeugt werden, stark gewachsen. Zusammen mit der Erfassung von Verwendungs-, Standort- und Bewegungsdaten aus mobilen Geräten kommt dabei der Erfassung der vielfältigen Nutzungen des Internets, wie der Kommunikation in sozialen Onlinemedien, Suchmaschinenabfragen, Webseitenbesuche und Auswertung von Browserhistorien, der Nutzung des Onlinehandels und weiterer Internetdienste (z. B. Streamingdienste) sowie elektronischer finanzieller Transaktionen und Bezahlsysteme eine besondere Rolle als Quelle personenbezogener Daten zu (Christl & Spiekermann 2016; Christl 2017; Constantiou & Kallinikos 2015; Weichert 2013; Pasquale 2015). Erfassung und kommerzielle Auswertung von personenbezogenen Daten werden oft als Gegenleistung für die „freie" Nutzung der Internetdienste wie Suchmaschinen oder soziale Onlinemedien erlangt („personal data as counterperformance") (EDPS 2017). Auch über die fortschreitenden Verbreitung sogenannter „smarter" Geräte („smart homes", „smart cars", „wearables", „fitnesstracker", „persönliche Assistenten" bzw. virtuelle Assistenten oder Sprachassistenten etc.) bzw. der Realisierung des „Internet der Dinge" mit der verbreiteten Ausstattung mit Sensoren und Vernetzung von Gegenständen werden sehr große Mengen an personenbezogenen Daten bei der Nutzung der Produkte und Dienste erfasst.

Die Ausweitung der Menge personenbezogener Daten hat mehrere für die Studie relevante **Konsequenzen** für Differenzierungen und Diskriminierungsrisiken. Durch die Ausweitung der Datenbasis werden viele Differenzierungen von Personen mit Verwendung von Algorithmen überhaupt erst ermöglicht. Insbesondere die Zusammenführung von Daten – entweder unternehmensintern oder mithilfe des Datenhandels bzw. des Daten-Brokerages – macht Differenzierungen auf Basis von umfassenden Personenprofilen möglich. Neue Formen der Datenanalyse, insbesondere weite Teile des maschinellen Lernens, funktionieren vor allem erst auf der Basis großer Datenmengen sinnvoll.

2.2.2 Ausweitung der algorithmenbasierten Analysemethoden

Im Folgenden werden einige für die Studie besonders relevante Entwicklungen, die teils über mehrere Jahrzehnte abgelaufen sind und auch weiterhin gegenwärtig stattfinden, skizziert. Dabei ist zu beachten, dass die verwendeten Begriffe sich teilweise mit einer großen Überlappung auf dieselben Entwicklungen beziehen oder Entwicklungen auf unterschiedlichen Darstellungsebenen beschreiben können, sie daher nicht trennscharf voneinander abzugrenzen sind.

Mit den Methoden des **Data-Minings** sollen Erkenntnisse bzw. statistische Zusammenhänge in großen Datensätzen ausfindig gemacht werden (Custers 2013; Calders & Custers 2013; Linoff & Berry 2011: 2). Kennzeichnend sind dabei automatisierte Verfahren, die Muster bzw. Regelmäßigkeiten in den Datensätzen aufdecken und bei denen, anders als bei klassischen statistischen Analysen, keine zu überprüfenden Hypothesen über mögliche Zusammenhänge zwischen Variablen als Grundlage des Vorgehens vorliegen. Die Ergebnisse, die als Menge von ermittelten Zusammenhängen erzeugt werden, werden auch als Modelle bezeichnet. Ergebnisse bzw. Modelle können für die Bildung von Klassen bzw. Kategorien genutzt werden, in denen Personen automatisiert zugeordnet werden können. Dabei erfolgt auch die Kategorienbildung in vielen Fällen allein aus der automatisierten Erzeugung von Korrelationen (Barocas & Selbst 2016: 677).[9]

Die Entwicklungen, die unter dem unscharfen Sammelbegriff **„Big Data“** gefasst werden, zielen vor allem ab auf die Zusammenführung und Verarbeitung von großen und unterschiedlichen Datensätzen, d. h. die in ihren Formaten heterogen sein können (z. B. Zahlen-, Bild-, Text-, Video-, Audioformate) und in unterschiedlichen Kontexten erfasst wurden. Algorithmen dienen hier vor allem der Datenauswertung. Werden personenbezogene Daten verwendet, leisten Algorithmen der Big-Data-Analytics unter anderem der Erstellung und Pflege umfangreicher Persönlichkeitsprofile Vorschub, die oft zur Vorhersage von Verhalten (z. B. des erwartbaren Kaufverhaltens) eingesetzt werden. Im engen, technischen Sinne dienen Big-Data-Techniken bzw. Big-Data-Analytics vor allem der automatisierten Erfassung, Aufbereitung, Verwaltung und Analyse von großen Daten-

[9] Siehe Näheres in Abschnitt 3.4, ab S. 31, und Abschnitt 5.1, ab S. 77.

mengen (Chen, Mao & Liu 2014). Im weiten Sinne umfasst der Begriff „Big Data“ auch die organisatorischen Vorkehrungen, Praktiken und Geschäftsmodelle, bei denen die verknüpfende Verarbeitung von großen, teils heterogenen Datenmengen die zentrale Rolle spielt und vor allem auf Prognosen und Reaktionen in Echtzeit abzielt (z. B. Zuboff 2015; siehe auch Kolany-Raiser u. a. 2018; Hoeren & Kolany-Raiser 2018).

Der Übergang zwischen Data-Mining und dem **maschinellen Lernen**, das üblicherweise als Teilgebiet der **künstlichen Intelligenz**[10] gesehen wird, ist fließend. Maschinelles Lernen ist ein unscharfer Sammelbegriff für sehr unterschiedliche Konzepte und Methoden, die sogar auch herkömmliche statistische Analysemethoden umfassen können.[11] Der Begriff maschinelles Lernen bezieht sich auf Verfahren zum automatisierten Auffinden von Korrelationen – auch als Zusammenhänge, Regelmäßigkeiten oder Muster bezeichnet – zwischen Variablen in einem Datensatz. Dabei versucht man, den menschlichen Prozess des Lernens nachzubilden, indem aus einer großen Zahl von Beispielen in Form von Lern- bzw. Trainingsdatensätzen maschinell (aber nicht ohne Involvierung von Menschen) die relevanten Muster bzw. Merkmale für das zu analysierende Objekt identifiziert und als Modell erzeugt werden. In den meisten Fällen dienen Verfahren des maschinellen Lernens zur Erzeugung von Vorhersagen oder Schätzungen von Ergebnissen (nach Lehr & Ohm 2017: 671). Zusammen mit der weiteren Steigerung der Leistungsfähigkeit und Kostensenkungen bei Rechnern sowie der verbreiteten Nutzung von Cloud-Rechenzentren hat vor allem das Anwachsen der Mengen an personenbezogenen Daten, die in ausreichender Qualität und Menge zur Verfügung stehen, dazu beigetragen, dass Systeme mit maschinellem Lernen gegenwärtig auf immer mehr Situationen mit Personen angewandt werden. Es existieren zahlreiche solcher Anwendungen, die auf die Mustererkennung in Datenbanken (z. B. Muster von betrügerischem Verhalten in Finanzdaten), die Bildverarbeitung („computer vision“) oder die Sprachverarbeitung in Text- oder Audioform, u. a. die Verarbeitung natürlicher Sprache („natural language processing“) abzielen. Bei der Bildverarbeitung sind vor allem Gesichtserkennungssysteme zu nennen, die nicht nur auf die Wiedererkennung von Personen, sondern auch

[10] Bisher hat sich keine allgemein akzeptierte Definition von künstlicher Intelligenz (KI) durchgesetzt. Viele Autor*innen beschreiben KI als technische Mittel, um menschliche Intelligenz nachzubilden.

[11] Siehe z. B. Domingos (2012), The Royal Society (2017), Jordan & Mitchell (2015), Alpaydin (2016), Leis u. a. (2018), Mullainathan & Spiess (2017), Strauß (2018), WIPO (2019).

auf die Erkennung von Zuständen (z. B. emotionale Zuständen, s. u.) und Verhaltensweisen von Personen ausgerichtet sind. Maschinelles Lernen überschneidet sich mit anderen Verfahren der Statistik und Datenverarbeitung wie dem Data-Mining. Die grundlegenden Algorithmen, die zu neuen Systemen mit KI kombiniert werden, sind teilweise seit Jahrzehnten bekannt (z. B. Regressionsmethoden), teilweise sind neue Lernalgorithmen hinzugekommen. Bei letzteren haben in den letzten Jahren die sogenannten „Künstlichen Neuronalen Netze", vor allem das „deep learning"[12], Beachtung gefunden.

Data-Mining, Big-Data-Analytics und maschinelles Lernen können zum **Profiling** eingesetzt werden. Der Begriff Profiling umfasst algorithmenbasierte Techniken und Praktiken der Verarbeitung großer Mengen an Daten, die der Erstellung und Verknüpfung, Aktualisierung und Verwendung von Datensätzen über natürliche Personen zur Erstellung eines umfassenden Bildes einer Person oder einer Gruppe von Personen dienen, das vor allem zur Kategorisierung, Bewertung, Prognose und Entscheidungsfindung herangezogen wird. Oft werden Profile erstellt, um von bekannten Merkmalen oder Personen auf Korrelationen beruhend Kategorien zu bilden, in die über Identifizierung gemeinsamer Merkmale auf die Kategorienzugehörigkeit unbekannter Personen geschlossen wird. Bekannteste Beispiele für den Einsatz von Profiling sind Gesetzesvollzug, Grenzkontrollen, kommerzielle und staatliche Überwachung der Internetnutzung („web tracking"), Marketing und Versicherungswesen (Hildebrandt & Gutwirth 2008; van Otterlo 2013; Helberger 2016; FRA 2018; Hänold 2018).[13]

---

[12] Deep-Learning-Methoden gehören zu den Verfahren der „Künstlichen Neuronalen Netze" und nutzen mehrere „Knotenschichten" von Berechnungsstufen („Neuronen"), die gewichtet miteinander verbunden sind, um Muster bzw. Korrelationen in einer Datenmenge (z. B. einem digitalen Bild eines Hundes) zu identifizieren. Dabei ist jeder Schicht dem Lernen mit einer anderen Stufe der Abstraktion gewidmet. Im Laufe des Trainingsprozesses wird die Gewichtung der Verknüpfungen angepasst. Im Beispiel des Hundebildes lernt die unterste Schicht etwa einfache Details wie z. B. die Werte eines Pixels, die nächst höhere versucht, Kanten zu lernen und höhere Ebenen lernen die Kombination von Kanten z. B. als Hundenase zu deuten. Vgl. Alpaydin (2016), Beck u. a. (2019).

[13] Eine Legaldefinition von Profiling liefert Art. 4 Nr. 4 DSGVO nach der Profiling „jede Art der automatisierten Verarbeitung personenbezogener Daten, die darin besteht, dass diese personenbezogenen Daten verwendet werden, um bestimmte persönliche Aspekte, die sich auf eine natürliche Person beziehen, zu bewerten, insbesondere um Aspekte bezüglich Arbeitsleistung, wirtschaftliche Lage, Gesundheit, persönliche Vorlieben, Interessen, Zuverlässigkeit, Verhalten, Aufenthaltsort oder Ortswechsel dieser natürlichen Person zu analysieren oder vorherzusagen […]" ist.

Unter **Scoring** wird die Zuordnung von Zahlenwerten zu Personen, meistens mit Zuordnung der Personen auf einer Skala und mittels der Berechnung von Wahrscheinlichkeitswerten für ein bestimmtes zukünftiges Verhalten verstanden (ULD & GP Forschungsgruppe 2014; Dixon & Gellman 2014; Weichert 2018; ausführlich dazu SVRV 2018). Anwendungen des Scorings, die derzeit besonders in der Aufmerksamkeit hinsichtlich möglicher ungerechter Behandlungen oder Diskriminierungen stehen, sind das Kreditscoring bei der Kreditvergabe, Scores der Arbeitsmarktchancen bei der Arbeitsvermittlung oder Risikoscores im Strafvollzug.

Zunehmend werden die genannten Verfahren der Datenverarbeitung für personenbezogene **Prognosen** verwendet. Im Gegensatz zu Ex-Post-Beurteilungen, bei denen überprüft wird, ob ein Sollwert oder Sollzustand eines Differenzierungsziels erreicht wurde oder nicht, sind Prognosen darauf ausgerichtet, Personen ex ante anhand von errechneten Wahrscheinlichkeiten in nach einem Differenzierungsziel gebildete Klassen einzuordnen oder einzelne Werte zuzuschreiben, die ausdrücken sollen, wie wahrscheinlich es ist, dass bestimmte Zustände in der Zukunft erreicht werden. In dieser Studie stehen vor allem Algorithmen zur Prognose im Fokus.

Insbesondere durch Entwicklungen des maschinellen Lernens als Teilgebiet der KI werden Analysen zur automatisierten **Identifizierung von Persönlichkeitsmerkmalen**, wie z. B. den Gesundheitszustand oder die sexuelle Orientierung, in einer zunehmenden Breite, mit (vermeintlich) besserer Genauigkeit, in (nahezu) Echtzeit und auf der Grundlage von bislang unüblichem Datenmaterial ermöglicht. Während beispielsweise die Erfassung des Persönlichkeitsmerkmals „Vertrauenswürdigkeit" bei der Bildung von Kreditscores lange vor allem auf der Zahlungshistorie und anderen finanziellen Informationen beruhte, werden heute Kreditscores auch mit Daten über die Kommunikation und Beziehungen in „sozialen" Onlinenetzwerken gebildet (Wei u. a. 2016).[14] Darüber hinaus lassen sich zahlreiche Beschreibungen und Experimente von Forschenden und Entwickelnden zur

[14] Eines der international führenden Unternehmen im Bereich Kreditscoring, Lenddo, verarbeitet nach eigenen Angaben neben Daten aus sozialen Onlinenetzwerken auch Telekommunikationsdaten, Browserdaten, Mobildaten, Daten von E-Commerce-Transaktionen und finanziellen Transaktionen, Daten aus der Analyse des Ausfüllens des Antrags sowie psychometrische Daten. Vgl. https://lenddo.com/ (zuletzt abgerufen am 17.4.2019). Siehe dazu auch das Patent, das das Unternehmen Facebook angemeldet hat, in dem ein Verfahren beschrieben wird, mit dem die Kreditwürdigkeit von Personen aus den Kreditscores der „Freunde" in dem Netzwerk gebildet wird (Meyer 2015).

Identifikation von Persönlichkeitsmerkmalen finden, hierzu gehören unter anderem:

- die Erkennung von emotionalen Zuständen anhand von Tastaturanschlägen (Epp, Lippold & Mandryk 2011),
- die Ableitung von sensiblen Informationen (u. a. Gesundheitszustand) aus Telefon-Metadaten (Mayer, Mutchler & Mitchell 2016),
- die Bestimmung von Naivität bzw. Raffinesse und der Nutzung bei Kreditangeboten (Ru & Schoar 2016),
- die Erkennung von Emotionen und Entwicklung psychodemografischer „Profile" auf Basis von Daten aus dem Onlinenetzwerk Twitter (Volkova & Bachrach 2015),
- das Erkennen von krimineller Neigung (Wu & Zhang 2016) und von genetischen Erkrankungen mit automatisierter Gesichtserkennung (Gurovich u. a. 2019),
- die Feststellung der sexuellen Orientierung anhand von Facebook-Kontaktlisten (Jernigan & Mistree 2009),
- die Ermittlung der „Rasse" bzw. ethnischen Herkunft anhand von Personenbildern (Fu, He & Hou 2014),
- die Ermittlung von psychologischen Eigenschaften (Extrovertiertheit, Introvertiertheit, Offenheit für Neuerungen) aus den „digitalen Fußabdrücken" wie „Likes"-Einträge oder Einträge bei dem Onlinenetzwerk Twitter (Matz u. a. 2017),
- die Ermittlung diverser Persönlichkeitsmerkmale, wie z. B. sexuelle Orientierung, Ethnizität, religiöse und politische Einstellungen, Alter, Geschlecht oder Intelligenz aus „Likes"-Einträgen bei Facebook (Kosinski, Stillwell & Graepel 2013) oder
- die Erkennung der sexuellen Orientierung, insbesondere Homosexualität, aus Personenbildern (Wang & Kosinski 2018).[15]

[15] Siehe auch Übersichten zu psychologischen Analysen auf Basis von Big-Data-Techniken Matz & Netzer (2017) oder zu Stimmungs- und Meinungsanalysen („sentiment analysis" oder „sentiment detection") Yue u. a. (2018).

Wie weit derartige Analysemethoden bereits in der Praxis angewandt werden, ist in den meisten Fällen unklar, da es keine systematische Erhebung dazu gibt. Was diese Beispiele jedoch erneut verdeutlichen, ist, dass es längst kein „belangloses" Datum mehr gibt – so schon 1983 das Bundesverfassungsgericht (BVerfG 1983) – und dass scheinbar „harmlose" Kommunikation und Verhalten potenziell die Basis und die Kriterien von Ungleichbehandlung und Diskriminierungen liefern können.

2.3 Algorithmen- und datenbasierte Differenzierungen

2.3.1 Typen der Differenzierung

Algorithmen werden innerhalb von Anwendungen der statistischen Datenauswertung, des Data-Minings, der Big-Data-Analytics und des maschinellen Lernens für die Differenzierung von Personen verwendet, entweder um neue Differenzierungsformen zu ermöglichen oder um bestehende Differenzierungen zu rationalisieren oder zu verfeinern.[16] Differenzierung von Personen meint dabei die Aufteilung einer Grundgesamtheit von Personen durch die Zuordnung von Personen zu Klassen, Kategorien, Gruppierungen bzw. (Markt-)Segmenten oder die Identifizierung von „Außenseitern". Gleichzeitig werden für die differenzierten Personengruppen oder Individuen verschiedene Objekte der Differenzierung unterschiedlich bereitgestellt oder durchgeführt. Eine Extremform der Differenzierung ist die Individualisierung, d. h. die Ausrichtung der Differenzierung auf ein einzelnes Individuum. Grundlegend müssen die Einzelpersonen oder Gruppierungen identifiziert werden, wozu die heute vielfältigen Möglichkeiten der Verarbeitung personenbezogener Daten und darauf beruhender Schlussfolgerungen genutzt werden (Gandy Jr. 2010: 30). Differenzierungen richten sich auf die personengruppen- oder individuenbezogene Bereitstellung von Informationen, Waren, Diensten, Entgelten, Positionen, Gewährung von Freiheiten etc. (siehe Tabelle 1).

[16] Vgl. auch Mittelstadt u. a. (2016). Siehe zu den geschäftlichen Anwendungen im Rahmen des Marketings bzw. der Gestaltung von Geschäftsbeziehungen z. B. Vercellis (2011).

Tabelle 1: Objekte der algorithmen- und datenbasierten Differenzierungen

Objekte	Beispiele
Informationen	Webseiteninhalte, Werbung, Suchergebnisse, (Partner-)Kontakte
Produkte und Dienste	Waren, einschließlich Informations- bzw. Medienprodukte, Immobilien, Versicherungen, Kredite, Ausbildung, medizinische Behandlungen, Infrastrukturdienste
Entgelte	Preise, Prämien, Tarife, Zinsen, Löhne, Lohnersatzleistungen
Entwicklungschancen und Positionen	Ausbildung, Arbeitspositionen, Arbeitsbedingungen, Ämter, sonstige Positionen
Freiheiten	(Nicht-)Einschränkungen (Inhaftierung, Kontrolle, Strafen)

Quelle: eigene Zusammenstellung

Differenzierungen und daraus folgende Ungleichbehandlungen von Personen können aus unterschiedlichen **Gründen** erfolgen: (1) Um unterschiedliche Risiken von Personen und Personengruppen zu handhaben, werden Risikoklassen oder Risikomaße (z. B. Risikoscores) gebildet (z. B. Rückfallrisiko bei Verbrechen, Kreditausfallrisiko, Risiko des Arbeitsplatzwechsels, Risiko ungeeigneter Stellenbesetzung). (2) Um den unterschiedlichen Wert der Kundschaft oder strategischen Wert von Personen in wirtschaftlichen Beziehungen zu bestimmen und zur Gewinnerzielung zu nutzen, werden Klassen und Maße des wirtschaftlichen Potenzials (z. B. Arbeitsproduktivität, Nachfrageverhalten, Umsatzpotenzial oder Zahlungsbereitschaft) gebildet. (3) Ebenso können Personen aus sozialen Gründen differenziert werden, beispielsweise aufgrund von Bedürftigkeit, Chancenverteilung oder Solidarität, wie dies bei besonderen Preisen für Studierende oder Rentner*innen sowie Förderprogrammen für bestimmte Personengruppen der Fall ist.

Algorithmenbasierte Differenzierungen können auch differenzierten **Verhaltenssteuerungen** dienen.[17] Dabei sind grundsätzlich „harte" und „weiche" Verhaltenssteuerungen zu unterscheiden. Die „harten" Verhaltenssteuerungen zielen auf den technischen Ausschluss von Regelabweichungen und die wirksame Regeldurchsetzung, wie z. B. durch programmierte Zugangs- und Nutzungsregeln bei Medienprodukten mit digitalem Rechtemanagement oder programmierten Vertragsbestandteilen bei Smart Contracts oder anderen Blockchain-Anwendungen ab. „Weiche" Verhaltenssteuerungen finden über finanzielle Anreize, bereitgestellte Informationen, Empfehlungen, sonstigen „nudges"[18] oder „dark pattern"[19] statt, die mit Algorithmen angeboten und administriert werden können. Dazu gehören beispielsweise Anreize durch personalisierte Versicherungstarife zu risiko-

[17] Siehe zur Diskussion, die unter verschiedenen Stichworten geführt wird, wie z. B. „Lex informatica" Reidenberg (1998), „Code is Law" Lessig (1999), (2006), „Regulation by Software" Grimmelmann (2005), „Regulation by Design" Yeung (2008), „Software as Governance" Shah & Kesan (2010), „Regulating Code" Brown & Marsden (2013), „Governing Algorithm" Barocas, Hood & Ziewitz (2013), „Governance by Algorithm" Just & Latzer (2016), „Algorithmic Regulation" Medina (2015), Yeung (2017), Hildebrandt (2018), „Governing through Technology" Kallinikos (2011), „Verhaltenssteuerung durch Algorithmen" Hoffmann-Riem (2017) oder „Software als Institution" Orwat u. a. (2010), Orwat & Bless (2016). Eine Übersicht über relevante Forschungsstränge findet sich z. B. in von Grafenstein u. a. (2018).

[18] Das sogenannte „nudging" (deutsch „Anstoßen", „Anschubsen") bezeichnet unternehmerische und politische Maßnahmen und Instrumente der Verhaltensbeeinflussung, die meistens auf Erkenntnissen der Verhaltensforschung beruhen. Mit ihnen wird versucht, an den Verhaltenseigenschaften der Menschen anzusetzen und ihre „Entscheidungs- und Auswahlarchitektur" vorzugeben, meist ohne ihre Wahlfreiheit einzugrenzen. Durch letzteres unterscheiden sie sich von Ge- oder Verboten. Sie zielen teils verdeckt auch auf unbewusste Verhaltenseigenschaften ab oder nutzen bestimmte menschliche Eigenschaften aus, wie den Hang, Mehraufwand zu vermeiden oder sozialen Normen und Erwartungen zu folgen. Beispiele sind Grundeinstellungen bei Computersystemen oder Onlinediensten, bestimmte Arten von Präsentationen oder Anordnungen von Waren (z. B. gesundes Essen vor ungesundem), Appelle, die soziale Normen ansprechen, oder bestimmte Darbietungen von Informationen auf Webseiten oder die Gestaltung von politischen oder unternehmerischen Maßnahmen als Spiel („gamification"). Die Abgrenzungen zu den länger bekannten finanziellen Anreizen oder Maßnahmen der Informationspolitik sind nicht immer klar und können diese auch enthalten. Vgl. Sunstein (2014), Smeddinck & Bornemann (2018), von Grafenstein u. a. (2018).

[19] Sogenannte „dark pattern" stellen Praktiken der Gestaltung der Elemente von Computersystemen, Onlinediensten, Plattformen des elektronischen Handels oder Webseiten an der Schnittstelle zu den Nutzenden dar, mit denen versucht wird, Nutzende zu nicht-intentionalem und für sie schädlichem Verhalten und Entscheidungen zu steuern. Inhaltlich besteht hier eine starke Überlappung zum „nudging", wobei „dark pattern" vor allem die manipulativen Praktiken meint, die zu Schäden bei den Nutzenden führen. Zu den Beispielen gehören überwachungsintensive Grundeinstellungen, Einschränkung der Wahl oder Erschwerung datenschutzfreundlicher Einstellungen, Zwang zu Registrierung, Verstecken von Kostenangaben oder Angaben, um Dringlichkeit der Entscheidung aufzubauen. Vgl. Forbrukerrådet (2018), Mathur u. a. (2019).

verminderndem Verhalten, die Selektion von Empfängern bestimmter Informationen oder die Gestaltung der Wahlmöglichkeiten der Nutzenden, wie z. B. personalisierte Anzeigen auf Webseiten oder Produktempfehlungen im Onlinehandel. Kennzeichnend für algorithmische Systeme ist, dass sie Verhaltenssteuerungen automatisiert in einem großen Maßstab ausführen, d. h. nicht nur für eine*n einzelne*n Nutzer*in, sondern für die gesamte Anzahl der relevanten Nutzenden (Yeung 2017, 2018: 19, 29).

Wirtschaftliche Differenzierungen mit der gruppen- oder individuenbezogenen Gestaltung von Produkten, Diensten und Entgelten haben eine lange Tradition in Marktwirtschaften. Dennoch werden die gesellschaftlichen Wirkungen und die Akzeptabilität von derartigen Differenzierungen immer wieder von neuem kontrovers diskutiert. Algorithmen- und datenbasierte Differenzierungen bringen diese Kontroversen über die Vor- und Nachteile von Differenzierungen wieder zu Tage. So wird zu den **Vorteilen** angeführt, dass differenzierte Informationen, Produkte oder Dienste besser den unterschiedlichen Präferenzen der Nachfrage entsprechen können. Dies kann nicht nur die Nachfrage, die Zufriedenheit der Kundschaft und die Identifikation von Personen der Kundschaft mit den Anbietenden und Gütern erhöhen, sondern auch die Kosten der Werbung durch Vermeidung „ungenutzter" Information bzw. Streuverluste bei der Werbung senken. Prinzipiell können auch niedrigere Preise für bestimmte Gruppierungen gewährt werden. Wirtschaftliche Differenzierungen sehen sich jedoch auch der Kritik hinsichtlich einer Reihe von **Nachteilen** ausgesetzt, insbesondere dass sie in unangemessener Weise die Zahlungsbereitschaften ausnutzen und einseitig die Konsumentenrente[20] abschöpfen. Differenzierungen, insbesondere in Form von Individualisierungen, können zudem die Wahlmöglichkeiten des Individuums und dadurch seine Autonomie einschränken (Barocas & Nissenbaum 2014: 54). Die Fragen, ob und wie aus Differenzierungen Diskriminierungsrisiken entstehen können, werden im weiteren Verlauf der Studie betrachtet.

[20] Die sogenannte Konsumentenrente ist die Differenz zwischen dem Preis, den eine nachfragende Person bereit wäre für ein Produkt oder ein Dienst zu zahlen, und dem Marktpreis, der sich tatsächlich für das Produkt oder den Dienst in dem Markt gebildet hat. Je größer die Differenz ist, desto größer ist der finanzielle Vorteil für die nachfragende Person. Mit der Differenzierung von Preisen wird u. a. angestrebt, höhere Preise bei denjenigen Nachfragenden durchzusetzen, die eine hohe Zahlungsbereitschaft für das Produkt oder den Dienst haben.

Insgesamt haben technische, methodische und organisatorische Entwicklungen dazu geführt, dass algorithmen- und datenbasierte Differenzierungen im Vergleich zu konventionellen Formen der Differenzierungen zu niedrigeren Kosten, in feinerem Detailgrad und vor allem entlang neuer Merkmale, wie z. B. vermeintlich ermittelte Charaktere und Persönlichkeitseigenschaften, erfolgen können (Agrawal, Gans & Goldfarb 2016, 2018). So haben die Entwicklungen der digitalen Technologien, einschließlich des Internets, dazu geführt, dass die Kosten zur Ermittlung und Nachverfolgung von Verhaltensweisen, Eigenschaften und Zuständen von Individuen („tracking“) reduziert wurden und eine bessere Verifikation der Identitäten möglich wurde (Goldfarb & Tucker 2017). Technisch dienen die vielfältigen Möglichkeiten des mehr oder weniger unbemerkten Trackings[21] (z. B. Besuch und Verhalten aus Webseiten, in sozialen Netzwerken und Onlinehandelsplattformen, beim Gebrauch von Apps) und die Registrierungen und Nutzerkonten (bzw. Accounts oder Logins) dazu. Dabei bedeutet Identifikation von Personen nicht nur das Wiedererkennen eines Individuums, sondern zunehmend auch die Identifikation der Merkmale und Zustände wie Alter, Gefühlszustände, sozialer Status oder sexuelle Orientierung. Bei Anwendungen mit kontinuierlichen Datenströmen (z. B. Beiträge in sozialen Onlinenetzwerken) und deren Auswertung sind zudem Differenzierungen mit ständigen Anpassungen und Experimenten bei den Geschäftsmodellen durch die Anbietenden möglich (Varian 2014).

Die gruppen- und individuenbezogene Differenzierung bzw. Personalisierung kann über computerbasierte Mittel und vor allem online besser als offline realisiert werden, da nicht nur die dazu notwendige Datenbasis identifizierter Personen bzw. Gruppen und deren Verhalten oder Zuständen vorliegt oder beschafft werden kann, sondern auch dadurch, dass die Realisierung der separaten gruppen- oder personenbezogenen Ansprache technisch besser umgesetzt werden kann. Denn auch die Anpassungskosten, wie z. B. die Preisanpassungskosten („menu costs“) oder die Kosten der auf Personen angepassten Informationsdarstellung, sind durch Informations- und Kommunikationstechnologien und Automatisierungen gesenkt worden (z. B. Varian, Farrell & Shapiro 2004: 12ff.). Personalisierung kann dann beispielweise auf Webseiten oder in Apps dadurch erreicht werden, dass die Betroffenen nur „ihr“ Angebot bzw. die auf sie bezogene Entscheidung wahrnehmen und nicht unmittelbar Vergleichsmöglichkeiten haben.

[21] Siehe dazu z. B. Klebert u. a. (2012).

Erst mithilfe von z. B. Vergleichsportalen oder im Austausch mit anderen Nutzenden kann die Vergleichbarkeit verbessert werden.

2.3.2 Anwendungsbereiche

Mittlerweile sind international vielfältige Anwendungsbereiche auszumachen, in denen Systeme mit algorithmen- und datenbasierten Differenzierungen eingesetzt werden, die Konsequenzen für die Lebensführung und die Entwicklungschancen von Personen haben.[22] In Kapitel 4 werden dazu Beispielfälle von Ungleichbehandlungen und Diskriminierungsrisiken genannt.

Im **Arbeitsleben** findet algorithmen- und datenbasierte Differenzierung über die Selektion von Stellenbewerbern und durch Bestimmung unterschiedlicher Entgelte und Arbeitsbedingungen für Arbeitnehmende statt. Die dazu verwendeten Systeme werden mit den Begriffen „talent analytics", „people analytics", „workplace analytics" oder „human resource analytics" benannt. Die erfassten personenbezogenen Daten umfassen Daten aus Bewerbungsunterlagen der Stellensuchenden, aus Arbeitsabläufen einschließlich der Kommunikation bis hin zu den Arbeitsergebnissen der Arbeitnehmenden. Neben der Bestimmung der Passfähigkeit von Bewerbenden zu der arbeitgebenden Organisation sollen die Systeme zur Überprüfung der Einhaltung von Arbeitsbestimmungen und Organisationsrichtlinien, der Produktivitätsbestimmung für Sanktionen oder Belohnungen bzw. Beförderungen sowie der Bestimmung der Wahrscheinlichkeit von Arbeitsausfällen oder Abwanderungen dienen.[23] Als Beispiel für „talent analytics" wird die digitale Aufzeichnung von Bewerbungsgesprächen genannt, die mit maschinellen Lernverfahren des „social sensing" ausgewertet werden (Chamorro-Premuzic u. a. 2017). Algorithmen der künstlichen Intelligenz werden zur Mustererkennung bei der maschinellen Durchsicht von digital vorliegenden Bewerbungsunterlagen, der Stimmen- und Wortwahlauswertung bei (elektronischen) Bewerbungsgesprächen oder zur Erkennung von bestimmten Gesichtsausdrücken (z. B. beim Lügen) bei Videobewerbungsgesprächen eingesetzt. Ein weiterer Bereich sind Onlineplattformen wie „soziale" Onlinenetzwerke oder Online-Ver-

[22] Übersichten liefern z. B. auch Lischka & Klingel (2017), Spielkamp (2019), Matzat u. a. (2019).

[23] Siehe dazu Rosenblat, Kneese & Boyd (2014), Burdon & Harpur (2014), Marler & Boudreau (2017), Chamorro-Premuzic u. a. (2016), 2017), Dzida (2017), Weichert (2018: 59–61), Angrave u. a. (2016), Kornwachs (2018), von Grafenstein u. a. (2018: 25–26).

mittlungsdienste, auf denen differenzierte Stellenanzeigen geschaltet werden können, die aber auch zum Management und zur Bewertung der vermittelten Leistungen durch die Leistungsnutzenden oder das Plattformunternehmen genutzt werden.[24]

Im **Handel** wird mit algorithmen- und datenbasierten Geschäftspraktiken versucht, neben zielgerichteten Produktempfehlungen und Werbung (z. B. Lecuyer u. a. 2015) auch Preisdifferenzierung in unterschiedlichen Formen zu erreichen.[25] Dabei lassen sich für die Preisdifferenzierung mit individualisierten Preisen für einzelne Individuen unter Verwendung von personenbezogenen Daten[26] nur wenige Beispiele bei einigen Onlineshops finden, häufiger ist die Gewährung von individuellen Preisnachlässen, Prämien oder Coupons bei Programmen zur Bindung der Kundschaft (z. B. Payback) (US CEA 2015; Schwaiger & Hufnagel 2018). Preisdifferenzierung kann in einer weiteren Form durch die Bildung von unterschiedlichen Versionen („versioning") eines Produktes oder Dienstes und von unterschiedlichen Marktsegmenten („market segmentation"), auf denen Produkte und Dienste je nach verschiedenen Abgabemenge, nach verschiedenen Zeitpunkten des Angebots oder in unterschiedlichen (Qualitäts- oder Ausstattungs-) Varianten zu unterschiedlichen Preisen angeboten werden, erfolgen. Üblicherweise ordnen sich die Nachfragenden bestimmten Marktsegmenten (Selbstselektion) zu oder können zwischen unterschiedlichen Versionen wählen.[27] Hier spielen algorithmische Verfahren vor allem eine Rolle bei der Datenanalyse zur Bildung von Markt- und Kundensegmenten, die auch mit anonymisierten Daten erfolgen kann. Bei einer dritten Form der Preisdifferenzierung verlangen die Anbietenden unterschiedliche Preise für unterschiedliche Gruppen (z. B. Seniorenrabatte).[28] Des Weiteren werden Algorithmen und Computersysteme im Management der Kundschaft zur Berechnung des sogenannten „customer lifetime value" eingesetzt, dem

[24] Siehe Übersicht z. B. bei Bogen & Rieke (2018).

[25] Zur Diskussion über die datenbasierte Preisdifferenzierung siehe US CEA (2015), Miller (2014), Ezrachi & Stucke (2016), Steppe (2017), Acquisti, Taylor & Wagman (2016), Zuiderveen & Poort (2017), Christl & Spiekermann (2016: 41ff.), Schwaiger & Hufnagel (2018), Zander-Hayat, Reisch & Steffen (2016), Tillmann & Vogt (2018a), (2018b). Zur Diskussion der Preisdifferenzierung nach Geschlecht und aus Sicht des Antidiskriminierungsrechts siehe an der Heiden & Wersig (2017).

[26] Auch als Preisdifferenzierung ersten Grades bekannt. Sie steht grundsätzlich in der Kritik, dass zu ihrer Realisierung personenbezogene Daten erforderlich sind und die Privatsphäre der Betroffenen beschnitten wird. Vgl. Varian, Farrell & Shapiro (2004: 14).

[27] Auch als Preisdifferenzierung zweiten Grades bekannt.

[28] Auch als Preisdifferenzierung des dritten Grades bezeichnet.

wirtschaftlichen Wert einer Person als Kundschaft über die gesamte (potenzielle) Lebensspanne der Kundschaftsbeziehung, um damit personen- oder gruppenbezogene Angebote und Werbung, zum Verhindern von Abwanderungen oder dem gezielten Beenden von Beziehungen („Demarketing") zu bestimmen (Blömeke & Clement 2009; Vercellis 2011).

In der **Kreditwirtschaft** werden neue Formen des Kreditscorings angewandt, bei denen die Datenbasis für die Scorebildung ausgeweitet sowie neue Analysemethoden angewandt werden. Während Risikoscores bei der Kreditvergabe seit langem verwandt werden, richtet sich die derzeitige Diskussion vor allem darauf, inwieweit der gegenwärtige Regulierungsrahmen die Risiken der neuen Verfahren ausreichend adressiert.[29] Im **Versicherungswesen** werden differenzierte Versicherungstarife angeboten, die auf neuen Methoden der Erfassung und Auswertung von personenbezogenen Daten beruhen. Dazu gehören die Telematiktarife für die Kraftfahrzeugversicherung mit der Erfassung und Auswertung des individuellen Fahrverhaltens.[30]

Im **Gesundheitsbereich** sind verhaltensbezogene Tarife der (privaten) Krankenversicherung zu nennen, bei denen die Erfassung von personenbezogenen Daten in Form von Bewegungsdaten oder Vitalparameter über Wearables oder Smartphones mit Apps erfolgt. In der Diskussion werden nicht nur die rechtlichen Bedenken vorgebracht, sondern auch ethische, wie negative Auswirkungen auf das Solidaritätsprinzip und Umverteilungseffekte.[31] Anwendungen mit algorithmischen Verfahren, u. a. des maschinellen Lernens, insbesondere zur Analyse von Bildmaterial, finden sich bei medizinischen Diagnoseverfahren. Ferner werden Systeme auch zur Zuordnung von Patient*innen zu bestimmten Behandlungen und Programmen eingesetzt.

Im **öffentlichen Bereich** werden Systeme im Bereich der Grenzkontrolle, vorausschauenden Polizeiarbeit („predictive policing")[32], zur Unterstützung von Gerichtsurteilen, bei der Überwachung von öffentlichen Räumen und

---

29 Siehe dazu Citron & Pasquale (2014), Weichert (2014), ULD & GP Forschungsgruppe (2014), Hurley & Adebayo (2016), Ferretti (2017), Wei u. a. (2016), Christl (2017), Eschholz (2017), Dorfleitner & Hornuf (2018).

30 Siehe dazu z. B. SVRV (2018), Hänold (2019).

31 Siehe dazu Weichert (2018), Deutscher Ethikrat (2017), ten Have (2013), Christl & Spiekermann (2016: 35ff.), Arentz & Rehm (2016), Bitter & Uphues (2017), Swedloff (2014), Selke u. a. (2018).

32 Siehe zu Predictive Policing z. B. Merz (2016), Robinson & Koepke (2016), Selbst (2017), Richardson, Schultz & Crawford (2019).

zur Identifikation von potenziellen Straftatbegehenden oder Terrorist*innen sowie zur Verwaltung von Sozialleistungen, Schulen und Hochschulen bzw. Studienplätzen eingesetzt.[33] Nach Matzat u. a. (2019: 28) sind diverse Systeme zur Entscheidungsunterstützung in Erprobung oder Vorbereitung für die Einführung in den Jobcentern der deutschen Arbeitsagentur.

Aus den genannten algorithmen- und datenbasierten Differenzierungen in den verschiedenen Lebensbereichen folgt nicht zwangsläufig, dass aus ihnen Diskriminierungen erfolgen. Im Kapitel 4 werden jedoch Beispielfälle von Ungleichbehandlungen und Diskriminierungen zu einzelnen Lebensbereichen gegeben.

2.3.3 Automatisierte Entscheidungen

Die Verwendung des Begriffs „automatisierte Entscheidung" ist in der wissenschaftlichen Diskussion und (Rechts-)Praxis mittlerweile üblich und thematisiert sowohl den Einsatz von Algorithmen zur Entscheidungsunterstützung von menschlichen Entscheidenden als auch die automatisierte Entscheidungsdurchführung, allerdings in nicht immer klarer Abgrenzung untereinander. Für beide Typen werden auch die Begriffe automatische Entscheidungssysteme bzw. „automated decision-making systems" (ADM Systems) oder „automated decision systems" verwendet (z. B. Zweig, Fischer & Lischka 2018; Zweig 2019).

Zur Veranschaulichung kann der Entscheidungsprozess abstrakt in mehrere Schritte unterteilt werden, die von der Aufnahme der Ergebnisse der Datenanalysen, der Bewertung der Situation und der Alternativen mit dem Abgleich mit vordefinierten Bedingungen, der Auswahl zwischen Alternativen bis hin zum Auslösen einer Aktion reichen.[34] Bei einer vollautomatisierten Entscheidung werden alle Schritte der Entscheidungsregeln durch Software ausgeführt. Dabei werden die Entscheidungsregeln durchaus von Menschen gesetzt oder bei maschinellen Lernverfahren Teile der Entschei-

[33] Übersicht z. B. in Spielkamp (2019).

[34] Siehe auch Parasuraman & Riley (1997: 232) (mit weiteren Nachweisen), die darauf hinweisen, dass Automatisierungen eher als spezifische Ausprägungen innerhalb eines Spektrums zwischen den Extremen der manuellen Handhabung einerseits und vollständiger Automatisierung andererseits, bei dem die Maschine alle Aspekte der Funktion kontrolliert, zu verstehen sind. Ähnlich auch Cummings (2004a), Vercellis (2011: 25–28). Zum Entscheidungsprozess vgl. ähnlich Kornwachs (2018: 174–179).

dungsregeln von Algorithmen auf Basis der Auswertung von Daten erzeugt. Beispiele für vollautomatisierte Entscheidungssysteme sind die automatisierte (Online-)Kreditvergabe, die Negativauswahl bei Systemen für das Management von Bewerbungen im Personalbereich, Empfehlungssysteme im Onlinehandel, automatisierte Preisanpassungen, Antrags- und Bearbeitungsverfahren im Versicherungswesen, Spamfilter bei E-Mail-Programmen oder (erwartete) automatisierte Verwaltungsakte bei Behörden, wie z. B. vollautomatisierte Steuerbescheide (Busch 2018; Weichert 2018; Straker & Niehoff 2018; Hänold 2019).

Wie Datenanalysen und Entscheidungsprozesse sind der Praxis zusammenhängen, ist sehr unterschiedlich und kann zur Veranschaulichung abstrakt in mehrere Typen unterteilt werden: (a) Die automatisierte Datenverarbeitung und der Entscheidungsprozess ist getrennt und die Ergebnisse der Datenverarbeitung werden quasi „per Hand" an automatisierte Entscheidungsprozesse übergeben bzw. dort als Entscheidungsregeln programmiert. Oder (b) die Datenverarbeitung ist in dem Entscheidungsprozess integriert. So können bei Verfahren des Data-Mining und des maschinellen Lernens die Ergebnisse in Form von optimierten Modellen als Programmteile direkt als Regeln der Differenzierung in Entscheidungssysteme eingebettet sein (Barocas & Selbst 2016: 677; Lehr & Ohm 2017; Kleinberg u. a. 2019). Die Unterscheidung in diese beiden Typen ist für die Erkennbarkeit von Diskriminierungen bedeutend, denn bei letzteren ist das Zustandekommen der Ergebnisse häufig weniger nachvollziehbar.[35] Für die weiteren Betrachtungen ist noch zusätzlich zu unterscheiden zwischen (a) statischen Systemen, die einmal oder in zeitlich getrennten Abständen Datenanalysen durchführen und die Entscheidungsregeln anpassen, und (b) dynamischen Systemen, die durch eine kontinuierliche Analyse von Datenströmen ständig die Entscheidungsregeln bzw. Modelle anpassen und optimieren (z. B. Yeung 2017).

Eine genaue Unterscheidung zwischen Entscheidungsunterstützung durch automatisierte Systeme der Datenverarbeitung und vollautomatisierter Entscheidungsdurchführung ist nicht nur aus ethischer Perspektive[36] hin-

[35] Siehe auch Abschnitt 3.4.

[36] So betonen Wiegerling, Nerurkar & Wadephul (2018) aus ethischer Perspektive, dass es sich bei solchen Entscheidungssystemen eigentlich nicht um ein „Entscheiden" des Systems handelt, denn solche Systeme kennen keine Folgenverantwortlichkeit und verfolgen keine eigenen Intentionen.

sichtlich der Zuschreibung von Verantwortung notwendig, sondern auch aus rechtlicher. Denn grundsätzlich besteht nach Art. 22 Abs. 1 DSGVO ein Verbot ausschließlich automatisierter Einzelentscheidungen, d. h. für diejenigen, die ohne menschliches Eingreifen erfolgen (siehe Abschnitt 6.2.3, ab S. 114, auch zu Ausnahmen und erlaubten Anwendungsformen). Zudem kann eine Tendenz in der Praxis bestehen, dass auch bei Systemen der Entscheidungsunterstützung menschliche Entscheidende, insbesondere aus arbeitsökonomischen Gründen, unterstellter Neutralität oder höherer Objektivität von Computerschlussfolgerungen, oder weil sie ein Abweichen von Computerempfehlungen nur schwierig gegenüber Vorgesetzten begründen können, die Computerempfehlungen direkt übernehmen und dass dadurch auch Systeme zur Entscheidungsunterstützung quasi den Charakter von Systemen der vollständig automatisierten Entscheidungsdurchführung erhalten.[37]

Die Vorteile von teil- und vollautomatisierten Entscheidungen werden in Effizienzgewinnen, Vermeidung von Fehlern und Vorurteilen bei menschlichen Entscheidungen sowie der Ermöglichung von Angeboten oder Reaktionen in „Echtzeit" gesehen. Neben Diskriminierungsrisiken werden als Risiken von automatisierten Entscheidungen potenzielle Manipulationen, Verantwortungsabwälzungen, mangelnde Nachvollziehbarkeit und Anfechtbarkeit durch Betroffene gesehen (Mittelstadt u. a. 2016; Busch 2018; Weichert 2018; Ernst 2017: 1027–1029; Zarsky 2016).

Beim **Vergleich zwischen menschlichen und automatisierten Entscheidungen** ist festzuhalten, dass viele menschliche Entscheidungen durch Vorurteile, Stereotypen oder sonstige Verzerrungen geprägt sein können. Hier besteht zunächst die Erwartung, dass automatisierte Entscheidungen über Computersysteme „neutraler" und „objektiver" sind, da Entscheidungsregeln ohne menschliche Emotionen oder subjektive Vorlieben vollzogen werden können, sowie Algorithmen wesentlich mehr Informationen in einer Entscheidungssituation verarbeiten können und damit „besser" entscheiden. Allerdings wird im Folgenden gezeigt werden, dass die Erwartungen an eine größere Rationalität und Neutralität nicht unbedingt erfüllt werden und dass sich durch teil- oder vollautomatisierte Entscheidungssysteme auch neue Risiken der Diskriminierung ergeben können.

[37] Ein ähnliches Phänomen ist der „automation bias", der dazu führt, dass Menschen den von Computern stammenden Antworten mehr vertrauen als ihren eigenen Einschätzungen, z. B. Cummings (2004a).

Eine weitere bedeutende Unterscheidung ergibt sich hinsichtlich der Anzahl der Entscheidungen. Haben menschliche Entscheidende, wie z. B. Sachbearbeiter*innen in einer Verwaltung oder einem Unternehmen, innerhalb von Entscheidungsregeln einen Entscheidungsspielraum, so können die dort vorkommenden Diskriminierungen eher punktuell, d. h. abhängig von der Anzahl der diskriminierenden Bearbeiter*innen, auf eine oder wenige Personen begrenzt sein.

Bei automatisierten Differenzierungsentscheidungen, die ein Diskriminierungspotenzial haben, weisen alle mit dem System getroffenen Entscheidungen das Diskriminierungsrisiko auf. Diskriminierungsrisiken können so zu einem Massenphänomen werden und leicht zu kumulierten Benachteiligungen führen.[38]

[38] Siehe Gandy Jr. (2010).

3. Diskriminierung

3.1 Begriffe und Verständnis

Welche Handlungen als diskriminierend gelten, wird in verschiedenen Gesellschaften, Zeiten und Regionen jeweils unterschiedlich angesehen. Die Abgrenzung ist das Ergebnis gesellschaftlicher Konflikte, Aushandlungen, Einigungen und Festschreibungen. Diese gesellschaftlichen Festschreibungen erfolgen vor allem in Menschenrechten und Grundrechten sowie den Gesetzen und Institutionen, die die Menschen- und Grundrechte konkretisieren und durchsetzen.

In dieser Studie wird dem in der EU und der Bundesrepublik üblichen Verständnis von Diskriminierung gefolgt und Diskriminierung als benachteiligende, ungerechtfertigte Ungleichbehandlung von Personen in Anknüpfung an ein geschütztes Merkmal verstanden.[39] Die Ungleichbehandlung erfolgt auf Grundlage der Kategorisierung und Zuschreibung von Merkmalen zu Personen. Die Kategorisierung und Bildung von Merkmalen kann beispielsweise auf Stereotypisierung, Vorurteilen oder auf rationalen Kalkülen beruhen, in Regel und Praktiken verborgen sein oder unabsichtlich erfolgen. In verschiedenen rechtlichen Katalogen werden Kategorien und Merkmale als geschützte Merkmale – synonym auch Diskriminierungsmerkmale genannt – festgeschrieben, nach denen Personen nicht in ungerechtfertigter Weise benachteiligt werden dürfen. Die wichtigsten sind in Tabelle 2 zusammengetragen. Ungerechtfertigt bedeutet vor allem, dass ein sachlicher Grund für die Ungleichbehandlung fehlt. Mit anderen Worten: Eine Ungleichbehandlung an sich kann auch in gesellschaftlicher Hinsicht akzeptabel sein, wenn ein anerkannter sachlicher Grund[40] dafür besteht (s. u.).

---

[39] Der Begriff „Diskriminierung“ wird hier wie im allgemeinen Sprachgebrauch im deutschsprachigen Raum sowie im europäischen und bundesdeutschen Recht mit seiner negativen Konnotation als gesellschaftlich unerwünschte Ungleichbehandlung oder Schlechterstellung von Personen verwendet, siehe dazu z. B. Berghahn u. a. (2016: 25). Dagegen wird häufig in der englischsprachigen Literatur, vor allem in wissenschaftlichen Abhandlungen, mit dem Begriff „discrimination“ jegliche Form der „Unterscheidung“ bzw. „Differenzierung“ von Personen bezeichnet, die durchaus auch gesellschaftlich vorteilhaft und akzeptiert sein kann.

[40] Vgl. Berghahn u. a. (2014: 57ff.), Schrader & Schubert (2018: AGG §3 Rn. 68ff., §§8, 9, 10 und 20).

Tabelle 2: Geschützte Merkmale[41]

Geschütztes Merkmal	Art. 3 GG	§§ 1 u. a. AGG**	Erwg. 71 DSGVO	Art. 9 DSGVO
„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

Quelle: eigene Zusammenstellung. *Nach Art. 9, Abs. 3 GG; **Darin besteht eine abgestufte Verwendung der Merkmale, z. B. gilt „politische Weltanschauung“ nicht im Zivilrechtsteil (siehe z. B. Wersig 2017).

Die im vorhergehenden Abschnitt 2.3 (S. 12ff.) genannten Objekte der Differenzierung (z. B. Waren, Immobilien, Positionen, etc.) kommen potenziell als Gegenstände der Diskriminierung und Diskriminierungsrisiken in Frage, sind aber nicht alle rechtlich geregelt. Differenzierung von Produkten und Diensten bedeutet immer auch, dass einer Person oder Gruppe etwas vorenthalten oder der Zugang erschwert werden kann, während einer anderen Person oder Gruppe dies erleichtert wird. Differenzierung der Preise bedeutet,

[41] Weitere Kataloge mit geschützten Merkmalen finden sich in der EU-Grundrechte-Charta (GRCh) und der europäische Menschenrechtskonvention (EMRK), die auch noch „Vermögen“ und „Geburt“ umfassen sowie die offene Klausel „sonstiger Status“ in der EMRK.

dass Personen der Zugang zu Ressourcen, Waren, Diensten, die zur wirtschaftlichen, sozialen und kulturellen Entfaltung der Persönlichkeit oder zum Aufbau von Fähigkeiten dienen, erschwert werden kann. Benachteiligungen drücken sich dann in konkreten ökonomischen Verlusten, wie z. B. bei Krediten oder Produkt- und Dienstleistungspreisen, oder Verwehrung von Zugängen zu Entfaltungsgelegenheiten oder Entwicklungschancen, wie bei Beschäftigungen, Wohnen oder Bildungsmöglichkeiten, aus. Selbst die Differenzierung von Informationen kann aus Antidiskriminierungssicht problematisch sein, wenn sich die Informationen auf Güter oder Positionen (z. B. Arbeitsstellen) beziehen, die der Persönlichkeitsentfaltung, sozialen Integration oder politischen Teilhabe dienen (z. B. Information- und Informationstechniken wie Internetzugang). Durch differenzierte Information bzw. Nichtinformation können nicht zuletzt die Wahlmöglichkeiten eingeschränkt werden.

3.2 Typen von Diskriminierung

Je nach dem Ziel von Untersuchungen, Diskursen, Diskussionen sowie Entscheidungen und Maßnahmen zur Antidiskriminierung werden verschiedene Typen von Diskriminierung unterschieden. Die häufigste ist dabei die Unterteilung in unmittelbare Diskriminierung und mittelbare Diskriminierung.[42] Nach § 3 Abs. 1 Allgemeines Gleichbehandlungsgesetz (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. Merkmale auf, die „Rasse"[43] oder ethnische Herkunft, Geschlecht, Religion oder Weltanschauung, Behinderung, Alter und sexuelle Identität umfassen (siehe auch Tabelle 2). Nach § 3 Abs. 2 AGG handelt es sich um **mittelbare Diskriminierung**, „[…], wenn dem Anschein nach neutrale Vorschriften, Kriterien oder Verfahren Personen wegen eines in § 1 genannten Grundes gegenüber

[42] Die unmittelbare Diskriminierung wird auch direkte Diskriminierung, im Englischen „disparate treatment" genannt, während die mittelbare Diskriminierung auch mit indirekter Diskriminierung, im Englischen „indirect discrimination", „systematic discrimination", „disparate impact" oder „unintended discrimination" bezeichnet wird.

[43] Im Folgenden wird der Begriff „Rasse" in Anführungsstriche 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 Biolog*innen gefolgt, die auf die fehlende wissenschaftliche Fundierung des Begriffs hinweisen. Siehe auch z. B. Wersig (2017: 42).

anderen Personen in besonderer Weise benachteiligen können, es sei denn, die betreffenden Vorschriften, Kriterien oder Verfahren sind durch ein rechtmäßiges Ziel sachlich gerechtfertigt und die Mittel sind zur Erreichung dieses Ziels angemessen und erforderlich."

Eine weitere Unterscheidung zwischen **präferenzbedingter** und **statistischer Diskriminierung** richtet sich auf unterschiedliche Motivationen der Entscheidenden, die eine Differenzierung vornehmen (Lorenz 1993). Bei der präferenzbedingten Diskriminierung („taste-based discrimination") beruhen Ungleichbehandlungen auf den persönlichen, auf Vorurteilen beruhenden Abneigungen bzw. Vorlieben der Entscheidenden gegen oder für eine bestimmte Personengruppe oder auf Abneigungen oder Vorlieben für bestimmte Produkte (Becker 1957/1971; zur Kritik siehe Arrow 1998). Präferenzbedingte Diskriminierung kann auf affektiver Zu- oder Abneigung beruhen. Daneben kann sie auch auf anderen Ursachen, wie sozialstaatliche Umverteilungsziele, basieren, etwa bei Höchstaltersgrenzen für Professor*innen, die nicht auf einem statistischen Nachweis beruhen, sondern eine Umverteilungsmaßnahme zugunsten jüngerer Professor*innen bzw. Anwärter*innen sind (Britz 2008: 23).

Diskriminierungsrisiken, die durch die Verwendung von Algorithmen und Datensätzen bei Differenzierungen von Personen erfolgen können, haben in vielen Fällen den Charakter von Risiken der statistischen Diskriminierung (Calders & Žliobaitė 2013: 53; Barocas & Selbst 2016: 677, 688–692; Goodman 2016; Williams, Brooks & Shmargad 2018).

3.3 Statistische Diskriminierung

Unter dem Begriff der statistischen Diskriminierung wird die ungerechtfertigte Ungleichbehandlung von Personen mithilfe von Ersatzinformationen verstanden.[44] Die Entscheidenden haben unvollständige Informationen über das Hauptmerkmal von Personen, über die eine Differenzierungsentscheidung getroffen werden soll. Bei der genauen Prüfung der Eigenschaften der

[44] Zur statistischen Diskriminierung siehe z. B. Britz (2008), Scherr (2016), Hellman (1998), Schauer (2003), Lippert-Rasmussen (2007), Gandy Jr. (2009), Fang & Moro (2011), Schauer (2018).

Person, um Informationen über das Hauptmerkmal der Differenzierung zu erlangen, entstehen Kosten und (zeitlicher) Aufwand. Kosten bzw. Aufwand werden von den Entscheidenden als so hoch beurteilt, dass sie auf Ersatzinformationen zurückgreifen, die vergleichsweise kostengünstiger bzw. mit geringerem Aufwand zu beschaffen sind.[45] Dabei kann es sich bei den Ersatzinformationen auch um Variablen einer Gruppenzugehörigkeit handeln, die die geschützten Merkmale sind (z. B. die Variable Alter) oder Variablen, die eine Korrelation zu geschützten Merkmalen aufweisen (z. B. Jahre an Berufserfahrung). Es können also zwei Typen von statistischer Diskriminierung unterschieden werden: (1) Wenn eine ungerechtfertigte Ungleichbehandlung vorliegt, bei der als Ersatzinformation ein oder mehrere rechtlich geschützte Merkmale herangezogen werden, kann von unmittelbarer statischer Diskriminierung gesprochen werden. Ein Beispiel wäre die Verwendung des Merkmals ethnische Herkunft, wenn ein vermeintlicher statistischer Zusammenhang zur Arbeitsproduktivität vermutet wird und Personen bestimmter ethnischer Herkunft von Arbeitsstellen ausgeschlossen werden. (2) Wenn Korrelationen von verwendeten, scheinbar neutralen Variablen zu geschützten Merkmalen bestehen, liegt eine mittelbare statistische Diskriminierung vor. Ein Beispiel wäre die Verwendung des Merkmals „Teilzeitbeschäftigung", bei dem jedoch eine Korrelation zum geschützten Merkmal Geschlecht besteht, da Frauen öfter in Teilzeit arbeiten. Bei beiden Typen werden die Ersatzinformationen auch als „Proxies" bezeichnet.

Werden als Ersatzinformation Variablen einer Gruppenzugehörigkeit verwendet (z. B. Alter von Arbeitnehmenden in Form einer Jahresgrenze), wird oft ein statistischer Zusammenhang zwischen diesen Variablen und dem Differenzierungsziel (z. B. Zuweisung in den Ruhestand von nicht mehr leistungsfähigen Arbeitnehmenden) sowie dem Hauptmerkmal der Differenzierung (Leistungsfähigkeit als Arbeitnehmender) angenommen. Dieser Zusammenhang wird dann generell für die weiteren Entscheidungen über andere oder alle individuellen Gruppenzugehörige angenommen, d. h., er

[45] Beispielsweise haben Personalentscheidende von Rechtsanwaltsbüros das Differenzierungsziel, das Risiko einer ungeeigneten Einstellung zu vermeiden. Das Hauptmerkmal, ein*e gute*r Rechtsanwält*in zu sein, ist jedoch nur schwer ermittelbar, weil (a) nicht unbedingt eindeutig ist, welche Eigenschaften und Qualitäten gute Rechtsanwalt*innen ausmachen, (b) einige eindeutig relevante Eigenschaften, wie z. B. die Urteilsfähigkeit, nur mit aufwendigen Tests ermittelt werden könnten und (c) selbst bei relevanten und testbaren Eigenschaften, wie z. B. der schriftlichen Ausdrucksfähigkeit, die Ermittlung kostspielige Beurteilungsverfahren erfordert. Daher ist es für die Personalentscheidenden effizienter, eine Ersatzgröße, wie z. B. dass Bewerbende zu den 10 Prozent Besten eines Jahrgangs einer angesehenen Universität gehören müssen, zu verwenden. Beispiel aus Hellman (1998).

wird generalisiert und die Ersatzinformation wird zur **Generalisierung** verwendet. Die Annahmen über den „statistischen" Zusammenhang können auch auf (vermeintlichem) Erfahrungswissen beruhen (Britz 2008: 8) oder auf statistischen Erhebungen und Nachweisen. Im weiteren Verlauf der Studie werden dieser „statistische" Zusammenhang und dessen Veränderungen mit Verwendung von Algorithmen näher betrachtet.

Nach Scherr (2016) liegt die Form der statistischen Diskriminierung auch vor, wenn Entscheidende auf Märkten (z. B. Arbeits- oder Wohnungsmarkt) zwar für sich beanspruchen, keine Vorurteile oder Diskriminierungsabsichten zu haben. Aber aufgrund einer (vermeintlich) unsicheren Informationsgrundlage über Eigenschaften, Fähigkeiten und Potenziale von sich individuell Bewerbenden werden stattdessen „[...] Annahmen über wahrscheinliche Unterschiede zwischen sozialen Gruppen, denen jeweilige Individuen zugerechnet werden, als Zusatzinformationen herangezogen, um den Entscheidungsprozess zu vereinfachen." (Scherr 2016: 5) (z. B. Geschlecht und Hautfarbe statt Qualifikationen). Dies wird oft getan, wenn der zeitliche Aufwand für die genaue Betrachtung des einzelnen Falles begrenzt ist: „In der Folge sind bereits mehr oder weniger plausible Annahmen über die wahrscheinlichen Eigenschaften kategorial unterschiedener Gruppen ein Einfallstor für Diskriminierung [...]" (Scherr 2016: 5).

Beispiel und Sonderfall

Ein aktueller Fall der statistischen Diskriminierung wird derzeit in Belgien untersucht. Dort verweigert der Energieversorger EDF Luminus die Stromversorgung an Personen, die innerhalb eines bestimmten Postleitzahlenbereichs wohnen. Für den Energieversorger stellt dieser Postleitzahlenbereich ein Gebiet mit vielen Personen mit schlechten Zahlungsgewohnheiten dar. Auch solvente potenzielle Nachfragende werden ohne Berücksichtigung ihrer individuellen Zahlungsfähigkeit von der Belieferung ausgeschlossen.[46] Dieser Fall stellt eine Sonderform der statistischen Diskriminierung dar, das sogenannte „redlining", die auf Basis der Ersatzvariable „Wohnort" erfolgt und ihren Namen durch das Einkreisen von Gebieten mit roten Linien erhalten hat (z. B. Barocas & Selbst 2016: 689).

[46] Auskunft durch Mitarbeiter*in der belgischen Antidiskriminierungsorganisation Unia – Interföderales Zentrum für Chancengleichheit, per E-Mail, 14. Nov. 2018.

Einige Autor*innen stellen heraus, dass es sich bei statistischer Diskriminierung um einen Diskriminierungstyp handelt, der auf „rationalen" Entscheidungen der Entscheidenden beruht. Daher wird diese Form auch oft der rationalen Diskriminierung zugeordnet (Gandy Jr. 2009, 2010; Hellman 2008). Im Gegensatz zur präferenzbedingten Diskriminierung haben dabei die Entscheidenden keine intrinsische Aversion gegen eine bestimmte Gruppe als solche, sondern die Diskriminierung erfolgt aus „rationalen" Kalkülen, um mit einem Informationsdefizit möglichst effizient umzugehen. Die Konzepte und Theorien der statistischen Diskriminierung sind in der Wirtschaftswissenschaft zunächst am Beispiel des Arbeitsmarktes entwickelt worden (Phelps 1972; Arrow 1973). Dieser Diskriminierungstyp ist später insbesondere auch für den Wohnungsmarkt, die Kredit- und Versicherungswirtschaft und verschiedene Ausprägungen der Altersdiskriminierung untersucht worden (z. B. Arrow 1998; Hinz & Ausprung 2017; Britz 2008).

Doch nicht alle Differenzierungen, die auf Kategorienbildung durch statistische Methoden und Auswertungen beruhen sowie Ersatzinformationen verwenden, sind Diskriminierungen im rechtlichen Sinne. Es ergeben sich Herausforderungen der Beurteilung ihrer Legitimität. Das Recht liefert Maßstäbe, nach denen beurteilt werden kann, ob eine Form der statistischen Differenzierung als ungerechtfertigt gilt. Hat sie beispielsweise die Eigenschaft einer mittelbaren Diskriminierung, sind nach dem AGG die sachliche Rechtfertigung und die Verhältnismäßigkeit zu prüfen.[47] (1) Die Verwendung des vermeintlich neutralen Merkmals (auch Verfahren, Vorschrift, Regel) kann durch ein rechtmäßiges Ziel sachlich gerechtfertigt sein, etwa aus arbeitsmarkt- und sozialpolitischen oder aus unternehmens- bzw. produktionsbezogenen Gründen. Dabei ist der alleinige Verweis auf Kostengründe nicht zulässig. Die sachliche Rechtfertigung muss immer in einer Einzelfallbetrachtung abgeschätzt werden. (2) Des Weiteren ist zu prüfen, ob die Mittel zur Erreichung des Ziels verhältnismäßig sind, d. h. erforderlich und angemessen. Dabei ist zu prüfen, ob kein milderes, ebenso geeignetes Mittel zur Verfügung steht und das Mittel zum angestrebten Ziel in einem angemessenen Verhältnis steht (hier nach Wersig 2017: 26f.) (Näheres siehe Abschnitt 6.1.3.2, S. 107ff.).

[47] Des Weiteren sind im AGG die Rechtfertigungsgründe für Ungleichbehandlungen an weiteren Stellen geregelt: Für die unmittelbare Diskriminierung bei Arbeitsverhältnissen in den §§ 5, 8, 9 und 10 AGG und für sonstige zivilrechtliche Verhältnisse in den §§ 5, 19 und 20 AGG sowie für die unmittelbare Diskriminierung in § 3 Abs. 2 AGG; nach Wersig (2017: 29–30).

3.4 Veränderungen bei der statistischen Diskriminierung

Bei Verfahren des Data-Minings, der Big-Data-Analysen und des maschinellen Lernens wird das Phänomen der statistischen Diskriminierung verändert, indem anstelle einer oder weniger Ersatzvariablen ganze Modelle treten, die eine Vielzahl von Variablen und deren gewichtete Relationen untereinander beinhalten. Solche Modelle werden durch die Auswertung von großen Datenmengen erzeugt und können als Entscheidungsregeln der Differenzierung bzw. Ungleichbehandlung in Software eingesetzt werden.

Im Allgemeinen ermöglichen Methoden des maschinellen Lernens, dass mathematische Modelle,[48] die die linearen oder nichtlinearen Zusammenhänge zwischen Variablen abbilden, an Datensätzen optimiert werden, die durch eine große, zuvor unbekannte Menge an relevanten Variablen gekennzeichnet sind. Beim praktischen Vorgehen werden häufig die Lern- bzw. Trainingsphase einerseits und die Anwendungs- bzw. Produktivphase andererseits unterschieden (Géron 2018). Als Ergebnis der Trainingsphase werden erkannte Zusammenhänge bzw. Muster als Modelle gespeichert und in der Anwendungsphase in Entscheidungsregeln eingebaut und auf neue Entscheidungssituationen übertragen.

Vereinfachend dargestellt besteht beim maschinellen Lernen das Trainieren eines Modells aus den folgenden Elementen: (1) Sammlung und Zusammenstellung eines Datensatzes; (2) Spezifizierung eines konkreten Ergebnisses, das im Datensatz vorhergesagt werden soll; (3) Entscheiden, welche möglichen Einflussvariablen gebildet und dem Trainingsalgorithmus zur Verfügung gestellt werden, um im finalen Modell berücksichtigt zu werden; (4) Konstruieren eines Verfahrens, um die beste Einflussvariable zu finden, die alle anderen Variablen benutzt, um das gewünschte Ergebnis vorherzusagen. Das Ergebnis ist das Differenzierungsmodell (Differenzierungsalgorithmus), das verwendet werden kann, um Vorhersagen über das Ergebnis, z. B. die Bewertung einer Person, zu machen; und schließlich (5) die Validierung des Verfahrens mit einem zurückgehaltenen Teil des

[48] In diesem Zusammenhang werden die Begriffe Modell und Algorithmus bzw. Lernalgorithmus teilweise synonym verwendet, z. B. Lehr & Ohm (2017). Der Anschaulichkeit halber wird im Folgenden aber möglichst nur von Modell gesprochen.

Datensatzes („hold out set“ oder Testdatensatz), der nicht für das Training verwendet wurde (Kleinberg u. a. 2019: 17f.).

In der Trainingsphase wird in einem iterativen Prozess ein mathematisches Modell mit Lernalgorithmen durch graduelles Anpassen der Parameter mithilfe des Einspeisens von Feedback optimiert, bis das Modell am besten an den Datensatz angepasst ist (bzw. „fittet“). Üblicherweise werden ein oder mehrere Ausgangsmodelle verwendet, an denen verschiedene Lernalgorithmen ausprobiert werden, bis einer der Lernalgorithmen die beste Leistung des Modells im Sinne der treffendsten Vorhersage oder Schätzung des Ergebnisses hervorbringt (Géron 2018: 30). In der Testphase wird das Modell auf den Testdatensatz angewandt und auf das sogenannte „overfitting“ oder „underfitting“ geprüft. Üblicherweise entstehen die Probleme des „overfitting“ oder „underfitting“, wenn das erzeugte Modell „zu sehr nur“ auf den Trainingsdatensatz „passt“, nicht gut verallgemeinert und nicht mehr gut zum ursprünglichen Datensatz „passt“, aus denen die Trainingsdaten gezogen worden sind.[49] Nach der Trainings- und Testphase kann das erzeugte Modell in der sogenannten Produktivphase verwendet werden, um auf Basis neuer Daten tatsächlich Vorhersagen oder Klassifikationen zu treffen.

Gegenüber dem herkömmlichen Programmieren kann die Verwendung von Verfahren des maschinellen Lernens Zeit- und Kostenersparnisse erbringen oder die Bearbeitung komplexer Aufgaben der Datenverarbeitung überhaupt erst ermöglichen. Mit herkömmlichem Programmieren müsste man, z. B. um E-Mail-Spam zu erkennen und auszufiltern, jeweils Regeln für einzelne Begriffe, Muster oder typische E-Mail-Bestandteile, von denen man weiß, dass sie häufig in Spam-E-Mails auftauchen, programmieren. Dies würde eine komplexe Liste erfordern, die zudem mit hohem Aufwand ständig neu programmiert werden müsste, wenn Spamversendende Begriffe oder Bestandteile verändern. Bei Verwendung des maschinellen Lernens für Spamfilter verwendet man zuvor von den Nutzenden als Spam gekennzeichnete E-Mails und lässt den Lernalgorithmus die relevanten Wörter oder Bestandteile erkennen (Géron 2018: 4–6). Das Beispiel zeigt, dass Verfahren des maschinellen Lernens insbesondere für die Aufgaben

[49] „Verallgemeinern“ bedeutet hier, dass das Modell aus einer gegebenen Anzahl von Trainingsbeispielen auf nie zuvor verwendete Daten verallgemeinert. „overfitting“ tritt dann auf, wenn das Modell angesichts der Trainingsdaten zu komplex ist und bei der Anwendung nicht mehr gut verallgemeinert. Von „underfitting“ spricht man, wenn das Modell nicht komplex genug ist, um komplexe Zusammenhänge in der Realität abbilden zu können. Vgl. Géron (2018: 17, 26–29).

geeignet sind, die für die direkte Programmierung „von Hand“ zu komplex bzw. zu aufwendig sind.

Zu den Charakteristika des maschinellen Lernens zählen, dass Verfahren des maschinellen Lernens mehr Dimensionen an Variablen verarbeiten können als konventionelle statistische Verfahren, sie teilweise ein höhere Genauigkeit bei der Vorhersage oder Kategorisierung erreichen können und dass üblicherweise viele Modelle vorliegen, die getestet und aus denen die passendsten ausgewählt werden können. Insbesondere dadurch, dass mehr Variablen für Differenzierungsentscheidungen zur Verfügung stehen, besteht zunächst Hoffnung, dass die Verwendung von geschützten Merkmalen als Kriterien der Differenzierung unattraktiv werden könnte und dadurch das Risiko unmittelbarer Diskriminierung verringern wird (US CEA 2015: 16).

Schon früh hat sich jedoch gezeigt, dass diese Charakteristika zwar als Vorteile angesehen werden können, aber auch eine Reihe von Nachteilen mit sich bringen. Beispielsweise ist eine höhere Genauigkeit mit einem Verlust der Einfachheit (und damit der Verständlichkeit) verbunden (Breiman 2001), oder es wird das Problem des „overfitting“ verursacht, sodass die maschinell erzeugten Kategorien nicht mehr denen entsprechen, die vom Anwender eigentlich angedacht waren (Hand 2006). Weitere Probleme, die zu Diskriminierungsrisiken führen können, werden in Kapitel 5 behandelt.

4. Beispielfälle von Ungleichbehandlungen, Diskriminierungen und Nachweismöglichkeiten

Im Folgenden werden Beispielfälle dargestellt, in denen Algorithmen innerhalb von Differenzierungsanwendungen zu Ungleichbehandlung von Personen geführt haben, die als potenzielle Diskriminierungsrisiken diskutiert werden oder die als Diskriminierung gerichtlich festgestellt wurden. Zudem verdeutlichen einige Beispiele die generellen Nachweismöglichkeiten von Diskriminierungen, ohne tatsächliche Diskriminierungen im rechtlichen Sinne nachzuweisen. Die Beispielfälle wurden im Rahmen einer Literaturrecherche zusammengetragen, die im Zeitraum Februar 2018 bis Juli 2019 durchgeführt und immer wieder aktualisiert wurde. Bei einer Vielzahl von Algorithmen zur Differenzierung und einer sehr großen, nahezu unüberschaubaren Menge an Anwendungen in diversen Systemen ist keine systematische Erfassung aller Anwendungen möglich. Daher ist zu beachten, dass die dargestellten Beispielfälle keiner vollständigen Systematik entsprechen können. Die Beispiele werden in Kapitel 5 möglichen Ursachenquellen von Diskriminierungsrisiken und gesellschaftlichen Konsequenzen sowie in Kapitel 6 Handlungsbedarfen und -optionen zugeordnet und diskutiert.

4.1 Arbeitsleben

Beispiel 1: Personalsoftware bei Amazon

Einem Medienbericht zufolge nutzte das Unternehmen Amazon seit 2014 ein in der Entwicklung befindliches Softwaresystem zur Suche und Bewertung von im Web aufzufindenden Lebensläufen potenzieller Mitarbeiter*innen. Das maschinelle Lernverfahren wurde an Lebensläufen trainiert, um nach Wortmustern zu suchen, die auf erfolgreiche Mitarbeitende schließen sollten. Während der Entwicklungszeit habe man bemerkt, dass das System nicht geschlechterneutral bewertet. Das System stufte Begriffe mit „women's“ und Namen von zwei ausschließlich weiblichen (Hoch-)Schulen herab. Als Trainingsdaten wurden Lebensläufe der letzten zehn Jahre

verwendet, die hauptsächlich von Männern stammten. Darin spiegelte sich die männliche Mehrheit an Beschäftigten in der Technologiebranche wider. Auch mit Anpassungen des Systems konnte man nicht sicherstellen, dass das System nicht andere Wege entwickelt hätte, mit denen Bewerbende diskriminiert worden wären. Das Entwicklerteam wurde 2017 aufgelöst. Personalangestellte hätten nach Angaben des Medienberichts die Empfehlungen des Systems in Erwägung gezogen, sich aber nicht vollständig auf das Ranking verlassen (Dastin 2018).

Beispiel 2: Onlineplattform TaskRabbit und Fiverr für Freiberufler*innen

In einer wissenschaftlichen Studie untersuchten Hannák u. a. (2016) die Onlineplattform für Freiberufler*innen TaskRabbit auf Ungleichbehandlungen. Der Onlinemarktplatz vermittelt kleinere Arbeitsleistungen, wie Arbeiten im Haushalt oder das Erledigen von Besorgungen. Zur Analyse wurden für den Zeitraum von fünf Jahren 3.707 Profile der anbietenden Personen von kleineren Arbeitsleistungen aus 30 Städten in den USA gesammelt und nach ihren Bewertungen („ratings"), der algorithmenbasierten Platzierung in der Rangfolge der Suchergebnisse der Onlineplattform und nach den Bewertungen durch Personen der Kundschaft in den Verhältnissen zu den Merkmalen Geschlecht und ethnische Herkunft mit Regressionsanalysen ausgewertet. Herausgefunden wurde, dass (1) Frauen, insbesondere mit Weißer Hautfarbe, zehn Prozent weniger Bewertungen als Männer mit vergleichbarer Qualifikation erhielten, (2) anbietende Schwarze Personen signifikant niedrigere Bewertungsscores erhielten als andere anbietende Personen mit ähnlichen Eigenschaften, und (3) der Algorithmus für die Rangfolge der Suchergebnisse signifikant mit den Merkmalen ethnische Herkunft und Geschlecht korrelierte, wobei die niedriger gestufte Gruppierung von Stadt zu Stadt unterschiedlich war. Die Forschenden empfahlen, dass Onlinemarktplätze proaktiv Verzerrungen identifizieren und vermindern sollen (Hannák u. a. 2016).

In einer ähnlichen wissenschaftlichen Studie (Hannák u. a. 2017) zu den Onlinemarktplätzen TaskRabbit und Fiverr mit 13.500 Profilen von ihre Arbeitsleistungen anbietenden Personen wurden ebenfalls Ungleichbehandlungen bei den Bewertungen der anbietenden Personen hinsichtlich der wahrgenommenen Merkmale Geschlecht und ethnische Herkunft („Rasse") nachgewiesen. Im Gegensatz zu TaskRabbit vermittelt der Onlinemarktplatz Fiverr kleinere „virtuelle" Arbeitsleistungen, wie z. B. Design von digitalen Dokumenten, Hilfe bei Programmierungen oder Videobear-

beitungen. Die automatisiert über Web-Crawling ermittelten Daten zu den Personen wurden durch menschliche Bewertende beurteilt und einem Geschlecht und einer ethnischen Herkunft zugeordnet, da diese konkreten Angaben ansonsten nicht auf den Plattformen genutzt werden. Die Bewertenden wurden über den Dienst Amazon Mechanical Turk beauftragt. Als Ergebnis wurde u. a. nachgewiesen, dass auf Fiverr Dienste anbietende Personen mit Schwarzer Hautfarbe weniger Bewertungen („reviews") und schlechtere Einstufungen („ratings") erhielten. Auch die Ausdrucksweise in den Bewertungen, die durch eine linguistische Analyse ausgewertet wurde, unterschied sich auf dem Dienst Fiverr nach Geschlecht und ethnischer Herkunft. Zusätzlich haben sie eine algorithmische Verzerrung beim Ranking der Suchergebnisse auf dem Dienst TaskRabbit herausgefunden, die sich in negativen Korrelationen zwischen dem Suchergebnisrang einerseits und Geschlecht und ethnischer Herkunft andererseits ausdrückten. Allerdings konnte für letzteres Ergebnis nicht die Ursache ermittelt werden. Stattdessen vermuten die Forschenden, dass der Algorithmus für die Suchergebnisse auf Basis der Bewertungen und Einstufungen der Nutzenden, die die Dienste in Anspruch genommen haben, gebildet wird. Da diese verzerrt waren, waren es auch die Rangfolgen der Suchergebnisse (ebd., S. 1915, 1927). Es kommt zu einer Fortsetzung der Diskriminierungsrisiken.

Beispiel 3: Geschlechterbezogene Ungleichbehandlung auf Onlineplattformen für Arbeitssuchende

In einer wissenschaftlichen Studie untersuchten Chen u. a. (2018) geschlechterbezogene Ungleichbehandlung bei spezialisierten Onlineplattformen bzw. Suchmaschinen für den Arbeitsbereich in den USA. Die Onlineplattformen bieten einerseits Arbeitssuchenden die Möglichkeit, auf der Seite Lebensläufe und Kurzprofile hochzuladen, und andererseits Personalsuchenden die automatische Sichtung und das Ranking von digitalen Lebensläufen. In der Studie wurden die Algorithmen zur Bildung der Rangfolge der Suchmaschinenergebnisse der Unternehmen Indeed, Monster und CareerBuilder betrachtet. Keine der Suchmaschinen gestattete, die Ergebnisse nach demografischen Merkmalen (z. B. Geschlecht, Ethnizität) zu filtern oder zu sortieren, sie erlaubten jedoch den Gebrauch von Ersatzvariablen, wie z. B. die Jahre an Berufserfahrung als Indikator für das Alter. Für die Erzeugung der Untersuchungsdaten wurden für 20 Städte und 35 Berufsbezeichnungen Suchen nach Bewerbenden mithilfe eines automatischen Webbrowsers durchgeführt, die zu Daten von 355.000 Bewerbenden führten. Die Forschenden haben das Geschlecht aus den Vornamen geschlussfolgert.

Als Ergebnis wurden Unterschiede nach Geschlecht in den Suchergebnissen bei allen drei Onlineplattformen herausgefunden. Bei der individuellen Fairness, die aus dem Rangplatz nach dem Geschlecht bei ansonsten gleichen Merkmalen bestimmt wurde, wurde eine leichte Schlechterstellung von Frauen nachgewiesen (allerdings mit nur geringen Effektstärken). Bei der Gruppenfairness, die vorläge, wenn der Rangfolgealgorithmus eine gleiche Verteilung der Ränge für Frauen und Männer zuordnen würde, wurden bei 12 von 35 Berufsgruppen Männer besser gestellt. Da die Webseiten keine Angaben zum Geschlecht erhoben, lag nach Einschätzung der Forschenden keine unmittelbare Diskriminierung vor, aber andere verdeckte Merkmale (Arbeitslosigkeit und Hochschule) könnten berücksichtigt worden sein.

Die Ergebnisse konnten durch die Forschenden nicht eindeutig interpretiert werden. Bei der individuellen Fairness könnte nach einer ihrer Hypothesen das Ergebnis auch dadurch zustande gekommen sein, dass der Algorithmus den Rang danach anpasst, wie viel Personalsuchende auf das jeweilige Profil geklickt haben. Die Ergebnisse zur Gruppenfairness lassen sich nach ihrer Einschätzung durch die ohnehin vorliegende strukturelle Ungleichheit in einigen der betrachteten Berufsgruppen (z. B. Softwareentwickelnde) erklären. Dadurch könne die Ursache von Ungleichbehandlungen nicht bei den Algorithmen gesehen werden, stattdessen schätzen die Forschenden die Suchmaschinen als erfolgreich ein, wenn es um Fairness im Sinne der gleichen Behandlung von gleichen Bewerbenden geht. Dies attestieren sie allerdings nicht für eine Fairnessinterpretation, die eine Verteilung der Arbeitnehmenden entsprechend der Verteilung in der Gesamtpopulation meint und eine aktive Einstellung von unterrepräsentierten Arbeitnehmenden bedeuten würde.

Beispiel 4: Diskriminierende Stellenanzeigen auf Facebook

Nach Angaben des Dänischen Instituts für Menschenrechte (Institut for menneskerettigheder) wird derzeit durch sie gerichtlich gegen Firmen vorgegangen, die die Differenzierungsmöglichkeit von Facebook genutzt haben bzw. nutzen, um selektive Stellenanzeigen nur für Männer auf der Onlineplattform zu schalten. Dabei richtet sich das Vorgehen nicht gegen Facebook selbst, sondern gegen die Anzeige schaltenden Firmen. Die Ungleichbehandlung wurde durch einen Journalisten aufgedeckt, der auch die Indizien recherchierte und dem Danish Institute for Human Rights zur Verfügung stellte. Weitergehende Informationen, z. B. über Algorithmen zum grundlegenden Profiling und zur Ermöglichung der selektiven Werbung, sind nicht bekannt. Auf der Grundlage des dänischen „Act on the Equal Treatment Board“ sowie des „Discrimination Act“ ist das Danish

Institute for Human Rights befähigt, ohne eine konkret klagende Person mögliche Diskriminierungsfälle vor das Tribunal zu bringen.[50]

Beispiel 5: Altersdiskriminierung durch Stellenanzeigen auf Facebook

Durch eine Untersuchung des Journalistenverbands ProPublica und der New York Times (Angwin, Scheiber & Tobin 2017) wurden Altersdiskriminierungen auf der Onlineplattform Facebook aufgedeckt. Dabei nutzten Unternehmen, wie u.a. Amazon, Verizon, UPS, Goldman Sachs und Facebook selbst, die Möglichkeit, Stellenanzeigen nur für bestimmte Altersgruppen auf Facebook zu schalten. Dies wurde durch die ca. 5.000 Optionen, Werbung zu personalisieren, dem sogenannten „microtargeting", ermöglicht. Die Einstellungsmöglichkeiten beruhen auf der detaillierten, algorithmenbasierten Auswertung der Daten über die Facebooknutzenden. Da die älteren Facebooknutzenden die Anzeigen nicht zu sehen bekamen, ergaben sich Fragen zur Ungleichbehandlung von Personen über 40 Jahren, die nach dem „Age Discrimination in Employment Act" (dem US-amerikanischen Gesetz gegen Altersdiskriminierung am Arbeitsplatz) verboten sind. Das Verbot bezieht sich auch auf „Hilfen" oder „Unterstützung" zur Altersdiskriminierung. Letztlich hat u.a. die Berichterstattung zu einem Gerichtsverfahren, angestrengt durch die Vereinigung „Communications Workers of America" u.a., vor dem District Court in San Francisco geführt.

Die Vereinigung „Communications Workers of America" u.a. haben dann eine Sammelklage gegen die Unternehmen T-Mobile, Amazon, Cox Communications und Cox Media in 2017 vorgebracht (United States District Court for the Northern District of California 2018). Aus der Verfahrensdokumentation geht hervor, dass die Klagenden den benannten Unternehmen und vielen weiteren Unternehmen vorwarfen, altersdiskriminierende Stellenanzeigen auf Facebook geschaltet zu haben. Zusammen mit anderen Gerichtsverfahren wurde als Ergebnis des Verfahrens eine Einigung mit Facebook erzielt (siehe Beispiel 7, S. 39).

Beispiel 6: Altersdiskriminierung durch Stellenanzeigen auf Onlinepersonalbörsen

Nach Angaben der Staatsanwaltschaft führte ein Verdacht auf Altersdiskriminierung zu einem formalen Vorgehen der Staatanwältin des US-Bundes-

[50] Angaben durch Mitarbeiter*in des Danish Institute for Human Rights per E-Mail an den Autor, März 2019.

staates Illinois, Lisa Madigan, gegen die Onlinepersonalbörsen Beyond.com, CareerBuilder, Indeed Inc., Ladders Inc., Monster Worldwide Inc. and Vault. In Schreiben an die Unternehmen sendete sie Warnungen, wonach durch den Zwang, dass von den Nutzenden bestimmte Alterserfordernisse auf den Webseiten eingehalten werden müssten, ältere Nutzende in der Jobsuche benachteiligt werden könnten. Beispielsweise gestatten die Unternehmen in den Webseitenmenüs Angaben zur Ausbildung und Berufserfahrung erst ab einer bestimmten Jahresgrenze oder die Informationen sind in Intervallen ab einem bestimmten Jahr anzugeben, die nicht für ältere Bewerbende mit früherer Ausbildung und längerer Berufserfahrung passen, sodass sie nur unvollständige Bewerbungsprofile erstellen können (Illinois Attorney General 2017). Algorithmen dienen hier der Steuerung, welche Informationen von Betroffenen angegeben werden können, wie die personenbezogene Datenanalyse und Klassifizierung erfolgen und sie ermöglichen die automatisierte und gezielte Adressierung bestimmter Personengruppierungen.

Beispiel 7: Diskriminierung nach Geschlecht durch Stellenanzeigen auf Facebook

2018 klagten die Bürgerrechtsorganisation American Civil Liberties Union (ACLU), Rechtsanwaltsfirma Outten & Golden LLP und die Gewerkschaft Communications Workers of America (CWA) zusammen mit der Equal Employment Opportunity Commission (EEOC) gegen das Unternehmen Facebook und zehn arbeitgebenden Unternehmen auf unrechtmäßige Diskriminierung nach dem Merkmal Geschlecht, da Stellenanzeigen auf Facebook nur an männliche Adressaten geschaltet wurden und dadurch alle Frauen und nicht-männliche Nutzende vom Erhalt der Anzeigen ausgeschlossen wurden.[51]

Neben den drei geschützten Merkmalen (Standort, Alter, Geschlecht), die Werbetreibende auswählen mussten, stellte Facebook zahlreiche weitere Kategorien für die detaillierte Bestimmung der zielgerichteten Adressierung zur Verfügung, die explizit oder implizit nach Geschlecht unterschieden, wie z. B. „[…] Single Dads, Single Moms, Soccer Mom, Working Moms,

[51] Siehe Informationen auf der Webseite von ACLU, https://www.aclu.org/cases/facebook-eeoc-complaints (zuletzt abgerufen am 28.8.2019), sowie in der Anklageschrift „Charge of discrimination", verfügbar auf der Webseite von ACLU: https://www.aclu.org/legal-document/facebook-eeoc-complaint-charge-discrimination (zuletzt abgerufen 28.8.2019).

Working Mother, Bad Moms, Strong Single Moms!, Proud Single Mother, The Single Moms Club."[52]

Zu den problematischen Praktiken, die in der Anklage angeführt wurden, wurde auch der sogenannte „lookalike audience"-Dienst gezählt. Bei diesem konnten Arbeitgebende oder Arbeitsvermittlungen Listen ihrer bestehenden Mitarbeitenden an Facebook übergeben. Facebook glich diese mit den Datensätzen über die Facebooknutzenden ab und stellte den Arbeitgebenden bzw. Arbeitsvermittlungen Listen mit demografisch ähnlichen Facebooknutzenden zur Verfügung, an die gezielt Stellenanzeigen gesandt werden konnten. Facebook nutzte bei der Verarbeitung Merkmale wie Standort, Alter, Geschlecht und Interessen. Dadurch liegt nach Meinung der Klagenden eine unmittelbare Diskriminierung vor.[53]

Im März 2019 wurde bei insgesamt fünf Gerichtsverfahren[54] eine Einigung zwischen dem Unternehmen Facebook und den klagenden Organisationen erzielt, bei der Facebook unter anderem versprach, einen separaten Bereich auf ihrer Plattform für die Anzeigen zu Personalstellen, Wohnungen und Krediten einzurichten, bei dem keine Adressierung nach Alter und Geschlecht sowie mit geschützten Merkmalen korrelierende Einstellungen mehr möglich sein sollen. Die zielgerichtete Werbeansprache auf Basis von Postleitzahlen oder innerhalb einer Region unterhalb eines 15-Meilen-Radius soll abgeschafft werden. Die Kategorien, die im „lookalike audience"-Dienst verwendet werden, sollen sich auf „country, region, profession and field of study" beschränken. Ebenso sollen dort Werbetreibende die Einhaltung von Antidiskriminierungsrechten bestätigen müssen. Das Unternehmen will ferner ein System der automatischen und menschlichen Überprüfung auf richtige Identifizierung und Einordnung der Werbungen einrichten.[55]

Beispiel 8: Ungleichbehandlung nach Geschlecht bei Berufsgruppenklassifizierung

Für die Online-Personalsuche sowie für automatisierte Verfahren der Personaleinstellung haben die Online-Präsenz bzw. Online-Biographie von Stel-

---

52 Siehe Anklageschrift in Fußnote 51, S. 13.

53 Siehe Anklageschrift in Fußnote 51, S. 11f.

54 Nach Gillum & Tobin (2019).

55 Sherwin & Bhandari (2019) und Informationen aus dem Dokument zur Einigung „Exhibit A – Programmatic Relief", verfügbar auf der Webseite von ACLU: https://www.aclu.org/legal-document/exhibit-describing-programmatic-relief-facebook-settlement (zuletzt abgerufen am 28.8.2019).

lensuchenden und Berufstätigen eine zunehmend wichtige Bedeutung, ob und wie sie gefunden werden und damit Zugang zu Beschäftigungspositionen erhalten. Automatische Entscheidungssysteme müssen dazu präzise die Beschäftigungen, Fähigkeiten, Interessen etc. erfassen können. Maschinelle Lernverfahren sollen dazu die Zuordnung von Personen zu Berufsgruppenklassifizierungen anhand der Beschreibungen (insbesondere die benutzten Worte und deren Kombinationen) in den Online-Präsenzen verbessern.

In einer wissenschaftlichen Untersuchung deckten De-Arteaga u. a. (2019) eine Ungleichbehandlung nach Geschlecht bei der Berufsgruppenklassifizierung auf, die auf existierende geschlechtsbezogene Ungleichheiten bei Beschäftigungen beruht. Zur Datenerfassung wurde die Suchmaschine Common Crawl benutzt, um 397.340 Online-Biographien zu sammeln. Sie testeten drei Verfahren des maschinellen Lernens zur semantischen Repräsentation mit dem Ergebnis, dass das Weglassen („scrubbing“) von expliziten Geschlechtsindikatoren, wie Vornamen oder Pronomen, nicht ausreichend war, um das geschlechtsbezogene Ungleichgewicht zu entfernen, und dass auch beim Fehlen der Geschlechtsindikatoren die Erkennungsrate („true positive rate“ bzw. Richtig-Positiv-Rate) mit den existierenden Geschlechterungleichgewichten in den Berufsgruppen korrelierte. Daher könnten Berufsklassifizierungen die Geschlechterungleichgewichte noch zusätzlich verstärken (De-Arteaga u. a. 2019: 2).

4.2 Immobilienmarkt

Beispiel 9: Diskriminierung bei Wohnungsanzeigen auf Facebook

Recherchen der Journalistenvereinigung ProPublica zeigten, dass das Unternehmen Facebook in seinen sozialen Netzwerken bei Anzeigen für Wohnungsvermietungen Diskriminierungen zuließ. Dazu hatte ProPublica selbst Anzeigen geschaltet und die Einstellungen für die zielgerichteten Anzeigen genutzt, damit diese nicht an Afroamerikaner*innen, Mütter mit High-School-Kindern, Personen, die eine Rollstuhlrampe benötigen, Personen jüdischen Glaubens, Auswandernde aus Argentinien und spanischsprechende Personen geschaltet werden. Diese Personengruppen sind nach dem US-amerikanischen Antidiskriminierungsgesetz für den Wohnungsmarkt, dem „Fair Housing Act“, geschützt. Alle Anzeigen wurden von Facebook genehmigt, obwohl die eigenen Unternehmensrichtlinien dies unterbinden sollten. Aufgrund der Recherchen wurde das U.S. Department of Housing and Urban Development (HUD), das US-amerikanische Ministerium für Wohnungsbau und Stadtentwicklung, welches unter anderem auch die Dis-

kriminierungsverbote im Wohnungsbau und Mietwesen überwacht, tätig (Angwin, Tobin & Varner 2017).

Im März 2018 wurde von der National Fair Housing Alliance (NFHA) Klage gegen Facebook erhoben. Im März 2019 wurde das Gerichtsverfahren mit einer Einigung beendet. Danach bietet NFHA dem Unternehmen ein Fair Housing Trainingsprogramm, zudem wird die Werbepolitik von Facebook regelmäßig kontrolliert und Facebook unterstützt Programme zur Ausweitung des Fair Housing. Des Weiteren versprach Facebook, ein separates Werbeportal für Wohnungs-, Beschäftigungs- und Kreditwerbung mit eingeschränkten Möglichkeiten der zielgerichteten Werbung einzurichten (NFHA 2019).[56]

Eine weitere Klage erfolgte im März 2019 durch das HUD (US HUD 2019b).[57] Die Anklageschrift richtet dabei unter anderem den Fokus auf die Rolle von Facebook bei der Selektion derjenigen Facebooknutzenden, die eine Anzeige gezeigt bekommen oder nicht. Diese Selektionsentscheidung würde weitestgehend auf Schlussfolgerungen und Vorhersagen über die Wahrscheinlichkeit getroffen, dass Nutzende auf die Anzeige reagieren. Die Schlussfolgerungen und Vorhersagen beruhen nach HUD auf einer Auswertung der Daten, die das Unternehmen über die Einzelperson hat, sowie auf Daten über andere Nutzende, die Facebook für ähnlich hält, und Daten über die „Freunde“ und andere mit der Person über Facebook verbundene Personen („associates“). Die Auswertung erfolge mithilfe von Verfahren des maschinellen Lernens oder anderen Vorhersagetechniken. Für die Selektionsentscheidung nutze Facebook das Merkmal Geschlecht und Proxies für andere geschützte Merkmale, wie z. B. die besuchten Webseiten, welche Apps ein*e Nutzer*in hat, wohin ein*e Nutzer*in während eines Tages geht und welche Einkäufe ein*e Nutze*n tätige. Diese Informationen würden auch für die Festlegung der Preise für die zielgerichtete Schaltung der Anzeige verwendet, die die Werbetreibenden zu entrichten haben. Dabei bestimme Facebook die Selektion derjenigen, die die Anzeige sehen, und nicht die Werbetreibenden. Zudem können Werbetreibende, die ein breites Publikum adressieren wollen, dies nicht erreichen, denn das Facebook-System entscheide ausschließlich selektiv nach den Charakteristika von Personen, die am wahrscheinlichsten auf die Anzeige reagieren (US HUD 2019a: 5).

[56] Siehe zur Einigung über die Anzeigen zu Stellenangeboten, Wohnungen und Krediten, die gleich mehrere Klagen betraf, Beispiel 7, S. 39).

[57] Stand März 2019.

Beispiel 10: Ungleichbehandlung nach ethnischer Herkunft auf Airbnb

In einer wissenschaftlichen Untersuchung weisen Edelman und Luca (2014) Ungleichbehandlungen nach ethnischer Herkunft auf dem kommerziellen Onlinemarktplatz für kurzfristige Vermietungen Airbnb.com nach. Für den Aufbau von Reputation und Vertrauen ermöglicht es Airbnb, dass Vermietende Selbstdarstellungen in Form persönlicher Profile platzieren, sowie dass Mietende Bewertungen über Vermietende und Vermietende über Mietende online angeben können. In einer Studie (2014) mit Auswertung von Daten zu Vermietungen in New York City, die einen Zusammenhang zwischen den Fotos der Vermietenden und den Mietpreisen herstellt, zeigen sie, dass Nicht-Schwarze („non-black") Vermietende zwölf Prozent höhere Mietpreise für vergleichbare Angebote erzielen können als Schwarze („black") Vermietende. Aus den Ergebnissen können sie allerdings nicht eindeutig schließen, ob es sich um eine Form der präferenzbedingten oder statistischen Diskriminierung handelt. Sie interpretieren die Ergebnisse als unintendierte Folgen der Mechanismen zum Aufbau von Reputation und Vertrauen.

In einer weiteren Studie (Edelman, Luca & Svirsky 2017) mit Daten zu Transaktionen über die Plattform Airbnb.com in den Städten Baltimore, Dallas, Los Angeles, St. Louis und Washington, D.C. weisen sie nach, dass Mietsuchende mit Namen, die nach Weißer Hautfarbe klingen, in 50 Prozent der Mietanfragen von den Vermietenden akzeptiert wurden, während Suchende, deren Namen nach Afroamerikanischer Herkunft („African American") klingen, nur in 42 Prozent der Mietanfragen akzeptiert wurden. Durch die Diskriminierung, d.h. die Ablehnung von Gästen, entstehen den Anbietenden „Kosten" in Form von entgangenen Gewinnen durch leer bleibende Zimmer. Nach Einschätzung der Autoren konnten mit der Untersuchung allerdings nicht die Mechanismen, die zur Diskriminierung führen, erklärt werden und auch diese Studie liefert keine eindeutigen Hinweise, ob es sich um präferenzbedingte oder statistische Diskriminierung handelt (Edelman, Luca & Svirsky 2017: 17).

Beispiel 11: Ungleichbehandlung nach ethnischer Herkunft auf Airbnb

Basierend auf der Studie von Edelman und Luca (siehe Beispiel 10), zeigen Gilheany u.a. (2015) beim Vermietungsdienst Airbnb, dass asiatische Vermietende für ähnliche Vermietungen 20 Prozent geringere Preise erzielen als Weiße Vermietende. Dazu haben sie insgesamt 101 als Weiße und „asiatisch" identifizierbare Airbnb-Vermietende in den Städten Oakland und

Berkeley untersucht. Die Ursachen der Unterschiede können die Autoren jedoch nicht eindeutig benennen. Auch bei diesem Beispiel zeigt sich, dass die Identifizierbarkeit der Anbietenden zum Risiko des diskriminierenden Verhaltens durch andere Nutzende führt. Es entstehen Fragen, ob Plattformbetreibende das Risiko mit ihren Möglichkeiten der Steuerung und Kontrolle, welche Informationen über die Plattformen kommuniziert werden, nicht besser verhindern könnten.

4.3 Handel

Beispiel 12: Nachweis von Preisdifferenzierung im Onlinehandel

In einer wissenschaftlichen Untersuchung von Onlinehandelsunternehmen im allgemeinen Handel und im Reisebereich nutzten Forschende (Hannák u. a. 2014) die Konten und Cookies von 300 fingierten Nutzenden, um Preisdifferenzierungen, d. h. auf einige Nutzende angepasste Preise, und Preissteuerungen, d. h. gesteuerte Anordnung der Suchergebnisse für besonders teure oder weniger teure Produkte auf hohen Platzierungen, nachzuweisen. Sie richteten Nutzerkonten ein, um die Auswirkungen verschiedener Merkmale der Nutzenden, wie die Art des Webbrowsers, Betriebssystems, das Vorhandensein eines Nutzerkontos oder die in der Vergangenheit gekauften und angeschauten Produkte herauszufinden. Zur Untersuchung haben sie mit Kontrollkonten tatsächliche Personalisierung von Störeinflüssen bzw. Rauschen getrennt. Zur Messung haben sie eine Metrik für die Informationsabfrage entwickelt. Für die Untersuchung wurden Arbeitsleistungen über die Crowdsourcing-Plattform Amazon Mechanical Turk eingekauft. Sie fanden bei neun der 16 untersuchten Onlinehändler Personalisierungen. Zwei Unternehmen nutzen Preisdifferenzierungen, um „Mitgliedern" reduzierte Preise zu gewähren. Zwei Unternehmen nutzen A/B Tests (zwei Varianten einer Webseite werden für unterschiedliche Personengruppen gezeigt), die eine Untergruppe von Nutzenden zu teureren Hotels führen, zwei Unternehmen nutzen personalisierte Suchergebnisse für mobile Geräte und ein anbietendes Unternehmen personalisiert die Suchergebnisse auf Basis vergangener Klicks und Käufe (Hannák u. a. 2014: 306). Das Beispiel zeigt zwar keine Diskriminierung im rechtlichen Sinne, da keine geschützte Personengruppe benachteiligt zu sein scheint, aber verdeutlicht, dass algorithmenbasierte Preisdifferenzierungen „von außen" ohne direkte Inspektion der Algorithmen nachgewiesen werden können.

Beispiel 13: Ungleichbehandlung in der Logistik bei Amazon

Einem Medienbericht zufolge berechnete bei Amazons Angebot „Same Day Delivery“ (Lieferung am Tag der Bestellung) ein Algorithmus die Gebiete, in denen das Unternehmen als erstes die damals neue Form der Belieferung einführte. Obwohl die ethnische Herkunft als Eingabe nicht von dem Algorithmus berücksichtigt wurde, wurden Stadtviertel ausgeschlossen, in denen überwiegend Menschen mit Schwarzer Hautfarbe lebten. Merkmale, die nach Angaben von Amazon in den Algorithmus berücksichtigt wurden, waren z. B. die Nähe zum nächsten Verteilzentrum und die Anzahl der Personen mit einer Prime-Mitgliedschaft in einem Gebiet. Vermutlich wurde eine Korrelation zum Merkmal der ethnischen Herkunft erzeugt. Der Fall wurde von Journalisten für sechs Städte recherchiert, die dazu die Verfügbarkeit des Dienstes jeweils nach Postleitzahlen testeten und die erhaltenen Angaben und Kartierungen mit Angaben der offiziellen Bevölkerungsstatistiken verglichen. Nach der Berichterstattung und Protesten versprach das Unternehmen, den Dienst auch in die zuvor benachteiligten Gebiete auszudehnen (Ingold & Soper 2016).

4.4 Werbung und Suchmaschinen

Beispiel 14: Stereotypen bei Ergebnissen von Suchmaschinen

Noble (2018) zeigte anhand von zahlreichen Beispielen, wie Suchmaschinen zur Verstärkung von Rassismus beitrugen, indem Suchergebnisse bei der Rangfolge der Ausgabe der Links, in den dargestellten Bildern oder bei Wortergänzungen bei der Auto-Vervollständigung im Sucheingabefeld auf herabsetzende Stereotypen verwiesen (z. B. überwiegend pornografische Bilder bei der Suche nach dem Stichwort „black girls“). Sie nutzte vor allem die Suchmaschine Google.

Ein frühes und viel zitiertes Beispiel wird in einem Medienbericht geschildert. Danach wurde in einer Foto-App des Unternehmens Google eine herabwürdigende Bezeichnung mit dem Tag „Gorillas“ bei Fotos von Schwarzen Personen automatisiert vergeben (Kasperkevic 2015).

Beispiel 15: Geschlechterunterschiede bei Suchergebnissen nach Bildern für Berufen

In mehreren wissenschaftlichen Studien untersuchten Kay u. a. (2015) die Ergebnisse von Suchen nach Bildern zu Berufsgruppen, die durch die Suchmaschine von Google ausgegeben wurden, und ob die Suchergebnisse Stereotypen in der Darstellung und in der Wahrnehmung verstärken. Die Un-

tersuchungen zeigten unter anderem eine Unterrepräsentation von Frauen in den Suchergebnissen zu denjenigen Berufsgruppen, die stereotypisch von Männern dominiert sind, im Vergleich zum Geschlechterverhältnis der offiziellen Beschäftigungsstatistik für diese Berufsgruppen. Dadurch komme es zu einer Verstärkung von Stereotypen. Des Weiteren wurde gezeigt, dass auch die Qualität der Darstellung (insbesondere im Sinne der dargestellten Professionalität), die durch Proband*innen bewertet wurden, für die Berufsgruppen höher ausfiel, die den Geschlechterstereotypen entsprachen. Ferner zeigten die Forschenden, dass die Wahrnehmung der Geschlechterverhältnisse bei den Suchergebnissen sich auf Vorstellungen über die tatsächlichen Geschlechterverhältnisse in den Berufen auswirken. Die Autor*innen diskutieren ihre Ergebnisse unter anderem hinsichtlich möglicher Verstärkungseffekte auf Ungleichverteilungen in Berufen, u. a. dass dadurch auch das Streben nach Karrieren in den Berufen beeinflusst oder begrenzt werden könnte. Sie erwarten jedoch Verbesserungen bei der automatisierten Kennzeichnung von Bildern (ebd., S. 3826).

Beispiel 16: Ungleichbehandlung bei zielgerichteter Werbung bei Gmail

Lecuyer u. a. (2015) zeigen anhand von Fallstudien, dass der Google-Dienst Gmail für die zielgerichtete Werbung auch sensible bzw. geschützte Merkmale benutzt (z. B. Gesundheit, religiöse Zugehörigkeit und Interessen, sexuelle Orientierung oder angespannte finanzielle Situation). Sie nutzten zur Untersuchung das von ihnen entwickelte Open-Source-System „Sunlight", das der Erkennung und dem statistischen Nachweis von Personalisierungen im Web, d. h. in Form von personalisierter Werbung, Empfehlungen oder personalisierten Inhalten, dient.

Beispiel 17: Ungleichbehandlung nach Geschlecht bei Werbung auf Google-Diensten

Eine weitere wissenschaftliche Untersuchung widmete sich den Einstellungsmöglichkeiten zur Werbung („ad settings") bei den Diensten des Unternehmens Google (Datta, Tschantz & Datta 2015). Für die Studie haben die Autoren das System „AdFisher" eingesetzt, das der Untersuchung der Wechselbeziehungen zwischen Nutzerverhalten, der Werbung von Google und den Einstellungen durch die Nutzenden dient. Das System sammelt mithilfe vom Computer simulierten Handelnden („agents") große Mengen von Personalisierungsergebnissen, wie z. B. personalisierte Werbeanzeigen, ein. Sie wiesen nach, dass Werbung für höherbezahlte Berufe vergleichsweise mehr an Männer als an Frauen gezeigt worden war, wenn das Geschlecht in den Einstellungen bei „Ad setting" auf Frauen geändert wurde,

obwohl das Verhalten beim Webbrowsing identisch war. Da die Untersuchung „von außen“ ein komplexes „Ökosystem“ der Onlinewerbung betrachtet hat, konnte die genaue Ursache der Ungleichbehandlung nicht aufgedeckt werden.[58]

Beispiel 18: Ungleichbehandlung nach ethnischer Herkunft bei Platzierung von Werbung

In einer empirischen Untersuchung weist Sweeney (2013) nach, dass Werbeanzeigen für kommerzielle Produkte der Dokumentation von Verhaftungen, Vorstrafen, Straftaten etc. („arrest records“) bei den Suchergebnissen der Suchmaschine von Google auf Webseiten mit dem Anzeigendienst unterschiedlich geschaltet werden, je nachdem, ob gesuchte Namen auf eine bestimmte ethnische Herkunft hindeuten. Derartige Dokumentationen werden beispielsweise von Arbeitgebenden bei der Bewertung von Arbeitssuchenden verwendet. In der Untersuchung erfolgten die Anzeigen für die Produkte öfter bei Suchen nach Namen, die auf eine nicht-weiße Hautfarbe hindeuteten, als für Namen, die nach Personen mit Weißer Hautfarbe klangen. Die Ursache wurde beim algorithmischen Verfahren gesehen, das Google mit dem Dienst Google AdSense verwendete, um gezielt Werbeanzeigen bei bestimmten Suchanfragen zu schalten. Dahinter stand ein automatisierter Echtzeit-Auktionsmechanismus, der Preise und Platzierung der Werbeanzeigen steuert. In diesen flossen u.a. die (bisherigen) Raten des Anklickens von Werbeanzeigen ein. Die Autorin konnte (noch) keine eindeutige Ursachen für die Ungleichbehandlung ausmachen (ebd., S. 52), aber u.a. werden über den Algorithmus für die Schaltung der Anzeigen das Verhalten von Nutzenden auf der Suchmaschinenwebseite, d.h. das erfasste Klickverhalten, sowie dessen vergangene und laufende Trends analysiert und dementsprechend gesellschaftliche Ungleichbehandlungen widergespiegelt.[59]

Beispiel 19: Ungleichbehandlung nach Geschlecht bei Werbung auf Facebook

Lambrecht und Tucker (2019) wiesen mit einer empirischen Falluntersuchung Geschlechterunterschiede bei Onlinewerbung für STEM (Science, Technology, Engineering and Math) Karrieren nach. Sie zeigten, dass die

[58] Zur weiteren Analyse des Falls, die zu ähnlichen Ergebnissen kommt, und rechtlichen Einordnung nach US-Recht siehe Datta u.a. (2018).

[59] Hinweise zur Erklärung werden in den Untersuchungen der nachfolgenden Beispiele 19 und 20 gegeben.

Testwerbeanzeige, die in 191 Ländern im „sozialen“ Onlinenetzwerk Facebook geschaltet wurde, 20 Prozent häufiger Männern als Frauen gezeigt wurde, obwohl Frauen eine höhere Wahrscheinlichkeit hatten, auf die Werbung zu reagieren („click through rates“). Die Autorinnen deuten dies so, dass die Gruppe „Frauen“ für die Werbetreibenden kostspieliger sind, da der algorithmische Mechanismus die Preise nach der Wahrscheinlichkeit, ob Nutzende die Werbung anschauen, d.h. nach der „click through rate“, berechnet. Die Autorinnen vermuten, dass Werbetreibende mehr Auktionen um die Zielgruppe „Frauen“ verlieren als um die Zielgruppe „Männer“. Eine vergleichbare ungleiche Darstellung der Anzeigen zwischen Frauen und Männern wurde auch für die Werbeanbietenden Google Display Network, Instagram und Twitter nachgewiesen. An den Ergebnissen verdeutlichen die Autorinnen, dass vor allem betriebswirtschaftliche Faktoren – in diesem Fall der Werbeauktionsmechanismus von Onlineplattformen – zu diskriminierenden Ergebnissen von algorithmenbasierten Differenzierungen führen können. Die Ergebnisse der Studie wurden durch die Untersuchung von Ali u.a. (2019) bestärkt (siehe Beispiel 20).

Beispiel 20: Ungleichbehandlung nach Geschlecht durch Algorithmus zur Werbeschaltung

Ali u.a. (2019) untersuchten die selektive bzw. personalisierte Werbung auf der Plattform Facebook auf mögliche Diskriminierungen. Dazu nutzten sie den Individualisierungsdienst „Custom audience“, den Facebook für Werbetreibende zur Verfügung stellte,[60] und bestimmten die Zielgruppe mit einer nach Telefonnummern präparierten Liste (und zur Kontrolle zufällig ausgewählte Nutzende). Sie testeten den Bilderkennungsalgorithmus und schalteten dazu verschiedene Anzeigen mit stereotypen Bildern für Frauen und Männer, wobei ein Teil der Anzeigen nur für maschinelle Bilderkennungsverfahren erkennbare „weibliche“ oder „männliche“ stereotype Merkmale aufwiesen (z.B. Baumaschinen oder Militär für Männer, Brautsträuße oder Parfüm für Frauen), während für menschliche Betrachtende diese

[60] Die Autor*innen gruppieren die verschiedenen Möglichkeiten der zielgerichteten Werbung auf der Facebook-Plattform in: (1) nach gruppenbezogene Selektion nach demografischen Merkmalen, wie Alter, Geschlecht, Standort oder Profilinformationen der Nutzenden, (2) nach einzelnen Personen, die gegenüber Facebook angegeben werden können, wie z.B. in Form von Listen von Namen, Adressen, Telefonnummern. Geburtstagen oder in Form von Personen, die durch Webtracking-Werkzeuge erkannt werden („custom audience“) oder (3) bestimmte Personengruppen können selektiv adressiert werden, indem auf ihnen gleichende Nutzende referiert wird, die zuvor schon ausgewählt wurden („lookalike audience“), vgl. Ali u.a. (2019: 4).

Bilder nicht zu erkennen blieben. Dennoch erreichten die präparierten Bilder mehrheitlich eine „weibliche“ oder „männliche“ Zielgruppe. Daher schlossen die Forschenden, dass Facebook die Bilddaten analysiert und automatisiert die Anzeigen entsprechend stereotypisch ungleich Geschlechtern zugeordnet und ausgeliefert hatte. Aus ihren Ergebnissen zogen sie die Schlussfolgerung, dass der unternehmensinterne Algorithmus festgelegt habe, welche Werbeanzeigen gegenüber welchen Nutzenden angezeigt worden waren, und dieser daher diskriminierend sein konnte. Der Algorithmus habe unabhängig von den Einstellungen agiert, die Werbetreibende in der Auswahl der Zielgruppen festlegen konnten. Die Autor*innen schränkten jedoch ein, dass ihre Ergebnisse nicht verallgemeinert werden können.

Die Ergebnisse der Studie verweisen auf mögliche rechtliche Konsequenzen, denn sie zeigen, dass nicht allein diejenigen, die die Werbeanzeigen schalten, verantwortlich sind, sondern auch das Plattformunternehmen Facebook, das die Algorithmen der Anzeigenregelungen entwickelt und einsetzt. Die Ergebnisse können auch in einen Zusammenhang mit der Klage des U.S. Department of Housing and Urban Development (HUD) bei Anzeigen im Wohnungsbereich gebracht werden (siehe Beispiel 9, S. 41) (Matsakis 2019).

4.5 Kreditwirtschaft

Beispiel 21: Diskriminierungen nach ethnischer Herkunft bei Auftreten von FinTechs

Mit einem Vergleich zwischen herkömmlicher Kreditvergabe und der mithilfe von Algorithmen zeigen Bartlett u.a. (2018), dass Diskriminierungen auf dem US-Kreditmarkt für Hypotheken aufgrund ethnischer Herkunft (Amerikaner*innen mit Afroamerikanischer und Lateinamerikanischer Herkunft) weiterhin stattgefunden haben, d.h. auch mit algorithmenbasierten Kreditentscheidungen. Durch Algorithmen und das Auftreten der FinTechs habe sich jedoch die Art der Diskriminierung verändert, nämlich von Diskriminierung durch menschliche Vorurteile oder Abneigungen hin zu illegitimen Anwendungen der statistischen Diskriminierung mit der Verwendung von Big-Data-Variablen. Mit Big-Data-Variablen werden gar nicht oder schlecht ermittelbare Variablen für die Risikobestimmung durch Ersatzvariablen ersetzt (z. B. nennen die Autor*innen High-School-Abschluss für Einkommenssteigerungen als eine von vielen Variablen). Zwar habe sich mit algorithmenbasierter Kreditvergabe der Wettbewerb erhöht, die Vergleichbarkeit zwischen Anbietern sei erleichtert worden und ebenso der Wechsel zwischen ihnen. Zudem diskriminieren FinTech-Unternehmen

nicht durch die Ablehnung von Krediten, wie dies konventionelle Kreditgeber tun, sondern durch höhere Preise bzw. Zinsraten. Bei letzterem war der Grad der Diskriminierung (gemessen an zusätzlichen Aufschlägen bei den Zinsen) jedoch genauso hoch wie bei der konventionellen Kreditvergabe.

Beispiel 22: Mehrfachdiskriminierung bei Onlinekreditvergabe durch Svea Ekonomi

In Finnland wurde, dem Urteilsdokument (YVTltk 2018) zufolge, im März 2018 ein Unternehmen der Kreditvergabe aufgrund der Verwendung einer ungeeigneten statistischen Methode, der Verwendung von geschützten Merkmalen und dem Auslassen einer Einzelfallprüfung der Zahlungsfähigkeit durch das Yhdenvertaisuus- ja tasa-arvolautakunta, dem Nationalen Nicht-Diskriminierungs und Gleichheitstribunal von Finnland, verurteilt (siehe zum Urteil YVTltk 2018).

Einem männlichen Antragsteller wurde eine Verlängerung eines Kredits, den er auf einer Webseite im Zusammenhang mit einem Onlinekauf beantragt hatte, durch das Kreditunternehmen Svea Ekonomi AB verweigert. Der Betroffene meldete den Fall dem Antidiskriminierungs-Ombudsmann, der ihn vor das Tribunal brachte. Das Tribunal entschied, dass das Kreditvergabeverfahren nicht mehr weiterverwendet werden darf und verhängte eine Strafe von 100.000 Euro. Das Tribunal begründete seine Entscheidung damit, dass ein Fall von direkter Mehrfachdiskriminierung vorgelegen habe, da die rechtlich geschützten Merkmale Geschlecht, Muttersprache, Alter und Wohnort verwendet wurden, und dass keine Einzelfallprüfung des Antragstellers hinsichtlich seines Kreditverhaltens und seiner Kreditwürdigkeit durchgeführt wurde, sondern stattdessen formale und abstrakte Kreditdaten, die auf dem Kreditverhalten anderer beruhen, verwendet wurden (YVTltk 2018). Bei diesem Fall zeigen sich die Charakteristika der statistischen Diskriminierung deutlich, d. h. das Vermeiden von individuellen Einzelfallprüfungen und stattdessen die Verwendung von Ersatzvariablen, die in dem Fall die geschützten Merkmale Geschlecht, Muttersprache, Alter und Wohnort waren und eine direkte bzw. unmittelbare statistische Diskriminierung bedeuten.

Es lag ein schematischer Kreditentscheid vor, der auf internen Daten des Kreditunternehmens, der Kreditdatendatei und auf Scoredaten beruhte. Der Score bildete u. a. die Faktoren Geschlecht, Sprache, Alter und Wohnort ab und basierte auf statistischen Zusammenhängen, die durch die Verwendung von Daten über andere Personen berechnet wurden. Danach wurden Männer häufigere Rückzahlungsprobleme zugeschrieben, weshalb sie eine

geringere Punktzahl als Frauen erhielten. Ebenso erhielten finnischsprachige Einwohner einen geringeren Scorewert gegenüber schwedischsprachigen Einwohnern. Der klagende Mann hat Finnisch als Muttersprache. Wenn er eine Frau gewesen wäre oder Schwedisch seine Muttersprache wäre, hätte der Score für den Kredit gereicht. Der Ombudsmann verwies darauf, dass der Kreditantragsteller keine Zahlungsausfälle hatte (YVTltk 2018: 7). Da der Mann zudem in einer Region wohnte, die durch das System mit einem Wert für unbekannte Gebiete belegt wurde, wurde darin ebenso eine Benachteiligung gesehen (YVTltk 2018: 6).

Der Ombudsmann unterschied in der Klage genau danach, dass das Kreditscoringsystem nicht dazu genutzt werden kann, präzise Informationen über die tatsächliche Situation einer individuellen antragstellenden Person zu erhalten, da das System nur eine statistische Bewertung darüber abgeben kann, wie wahrscheinlich es im Durchschnitt ist, dass Antragstellende dem Profil von Antragstellenden mit schlechtem Kreditscore entsprechen. Er wies zudem darauf hin, dass der Antragsteller nicht wie ein Individuum behandelt wurde, sondern wie ein „representative of statistical profiling", das hauptsächlich auf Variablen der geschützten Merkmale basiert, die der Kreditgeber auf alle Personen anwendet, die dem Profil entsprechen, wie Männer, die in einer bestimmten Wohngegend leben oder eine bestimmte Muttersprache haben und in einem bestimmten Alter seien (YVTltk 2018: 4f.). Die Methode führe dazu, dass Personen mit einem stabilen Einkommen und mit Anzeichen, die die Fähigkeit belegen, den Kredit zurückzuzahlen, keinen Kredit bekämen (YVTltk 2018: 5).

Das beklagte Kreditunternehmen wies in seiner Antwort darauf hin, dass nach dem Antidiskriminierungsrecht eine Ungleichbehandlung keine Diskriminierung darstelle, wenn die Behandlung auf einem Gesetz beruht und ein im Übrigen akzeptables Ziel hat und die Maßnahmen, um das Ziel zu erreichen, verhältnismäßig sind (YVTltk 2018: 8). Zudem hätte das Kreditverfahren dem Kreditdatengesetz (Credit Data Law) entsprochen und die Aufsichtsbehörde keine Beanstandungen gegen das Verfahren gehabt. Das Kreditverfahren sei ein Element von Online-Verkaufssystemen anderer Betreibender und diese Art von Online-Finanzierung sei an den Einkauf gebunden und erfolge schnell und automatisiert. Die individuelle Prüfung der Kreditwürdigkeit von kreditsuchenden Personen mit der Nutzung von persönlichen Informationen und Dokumenten, wie z. B. Gehalts- oder Steuerzertifikate, sei nicht geeignet für diesen Typ von Finanzierungsverfahren (YVTltk 2018: 10f.).

Das Tribunal entschied dennoch, dass das Verfahren nicht verhältnismäßig sei und deshalb nach den geltenden Diskriminierungsgesetzen nicht akzeptabel. Das Tribunal entschied ebenso, dass das Kreditunternehmen die Vermutung über Diskriminierung nicht widerlegen konnte (YVTltk 2018: 2). In dem Urteil wird auch festgestellt, dass die Abschätzungen der Zahlungsfähigkeit zunehmend auf Basis von Annahmen getroffen werden, die mit Daten erzeugt werden, die über andere Menschen gesammelt werden. Diese Annahmen können nicht dazu genutzt werden, dem Kreditantragsteller akzeptable Begründungen für die Kreditablehnung zu liefern, insbesondere dann nicht, wenn dem Kreditantragsteller keine Gelegenheit gegeben wird, seine tatsächliche Zahlungsfähigkeit und die beeinflussenden Faktoren klarzustellen (YVTltk 2018: 17).

4.6 Medizin

Beispiel 23: Verzerrte Datensätze bei einem Diagnosesystem

Bei einem System des maschinellen Lernens im Gesundheitsbereich, das das Risiko der Sterblichkeit für Patient*innen mit Lungenentzündung vorhersagen sollte, wurden verzerrte Trainingsdaten festgestellt. Das System sollte Entscheidungen, ob Patient*innen ambulant oder stationär behandelt werden können, unterstützen. Das auf einem neuronalen Netz basierende System ermittelte eine höhere Überlebenschance für Patient*innen mit Asthma, was jedoch gegen das medizinische Erfahrungswissen sprach. Die Verzerrung im Datensatz ergab sich dadurch, dass Patient*innen mit Lungenentzündung und einem (langfristigen) Asthmaleiden direkt auf Stationen der Intensivpflege gebracht worden sind. Dadurch wurden für Patient*innen mit Lungenentzündung und Asthma bessere Ergebnisse erzielt, als für Patient*innen, die „nur“ eine Lungenentzündung hatten. Obwohl das System korrekte Vorhersagen traf, entstand das Problem, weil diese Informationen über den Kontext nicht in das Entscheidungsunterstützungssystem einbezogen wurden (Caruana u. a. 2015). Der Fall verdeutlicht zudem die Probleme, die entstehen, wenn relevante aber nicht in Datensätzen repräsentierte Kontextfaktoren bei Entscheidungen unbeachtet bleiben (Cabitza, Rasoini & Gensini 2017: E1).

Beispiel 24: Diskriminierung nach ethnischer Herkunft bei Patientenzuordnung

Obermeyer und Mullainathan (2019)[61] beschreiben, dass sie Diskriminierung nach ethnischer Herkunft bzw. rassistische Diskriminierung bei einem weitverbreiteten, kommerziellen System für die Zuordnung von Patient*innen, die eine intensive medizinische Betreuung benötigen, zu einem „care management"-Programm nachgewiesen hatten. Die Zuordnung zum „care management"-Programm ist mit einer höheren Ausstattung mit Ressourcen verbunden. Dabei hatten Weiße Patient*innen eine höhere Wahrscheinlichkeit, dem Programm zugeordnet zu werden, als Schwarze Patient*innen in einem vergleichbaren Gesundheitszustand. Die Zuordnung erfolgte mithilfe eines algorithmisch erzeugten Risikoscores. In die Berechnung gingen Daten über die gesamten medizinischen Ausgaben eines Jahres sowie feingranulare Daten zur Inanspruchnahme von Gesundheitsleistungen des Vorjahres ein. Der Score gab demnach nicht den zu erwartenden Gesundheitszustand wider, sondern sagte die Kosten von Behandlungen voraus. Aus der Sicht der Autoren sind diese Vorhersagen über die Kosten auch genau und unverzerrt.

Sie sehen das Problem jedoch darin, dass Behandlungskosten zwar Proxies für den Gesundheitszustand darstellen können, aber diese unzureichende Ersatzinformationen sind. Denn auch andere Faktoren als allein der Gesundheitszustand bestimmt die Kostenhöhe, wie z. B. die ethnische Herkunft. So würden Schwarze Patient*innen abhängig von ihrem Gesundheitszustand weniger in der Behandlung kosten. Ein Algorithmus, der korrekt die Kosten für einzelne ethnische Gruppierungen vorhersagt, liefert notwendigerweise verzerrte Vorhersagen über die Gesundheitszustände. Sie führen das Problem auf die Bestimmung der Zielfunktion und die Auswahl der Label zurück, die einseitig auf die Optimierung von Kosten ausgerichtet seien und Externalitäten im Hinblick auf Gesundheit erzeugen (Obermeyer & Mullainathan 2019).

[61] Es liegt allerdings bisher nur ein Abstract vor.

4.7 Verkehr

Beispiel 25: Quasi-Segregation bei Navigationsdiensten

Der Navigationsdienst „Ghetto Tracker", bzw. nach Umbenennung „Good Part of Town" in den USA, der im Verdacht stand, auf „unsichere" Gebiete mit überwiegend nicht-weißen Bewohner*innen hinzuweisen, wurde nach Protesten wieder geschlossen. Allerdings verfügen nach Angaben von Silver auch verschiedene andere Navigationsdienste über ähnliche Funktionalitäten, mit denen vor „unsicheren" Nachbarschaften gewarnt wird (Silver 2013). Routenplanung ist eines der prominentesten Beispiele für den Einsatz von modernen Algorithmen, wobei hier vor allem auch die verarbeiteten Datensätze mit den diskriminierenden Bewertungen und Stereotypen die Ursache von Diskriminierungsrisiken sein dürften.

Beispiel 26: Möglichkeit der Ungleichbehandlung beim Fahrdienstvermittler Uber

Anhand der Fallstudie des Vermittlungsdienstes für Fahrdienstleistungen Uber zeigen Rosenblat u.a. (2017), dass das System der Fahrendenbewertung durch Kund*innen bzw. Nutzende des Fahrdienstes Quelle von Diskriminierungen nach ethnischer Herkunft (bzw. „Rasse") sein kann. Die Fallstudie beruht auf qualitativer Feldforschung mit ethnografischen Studien und Interviews der Fahrzeugführenden. Das Bewertungssystem ist ein grundlegendes Element des Unternehmens zur Qualitätssicherung der individualisierten, verstreuten Belegschaft. Die Bewertungen durch die Nutzenden sind ein Schlüsselfaktor in den automatisierten Bewertungen und Personalentscheidungen des Unternehmens, z.B. über „Aktivierung" der Fahrenden oder ihre „Deaktivierung" mit zeitweiliger Aussetzung oder vollständiger Beendigung des Verhältnisses. Sie wirken sich ebenso direkt auf die Verdienste der Fahrenden und die Chancen, höher bezahlte Arbeit zu bekommen, aus.

Die Autor*innen vermuten, dass Bewertungen durch Nutzende bzw. Kund*innen systematisch durch Ungleichbehandlungen nach ethnischer Herkunft und Geschlecht geprägt sind und verweisen dazu auf Erkenntnisse aus analogen Bereichen. Für den genauen Nachweis fehlte ihnen jedoch der Zugang zu den Unternehmensdaten über die Bewertungen der Nutzenden und die Zusammensetzung der Gruppe der Fahrenden. Dadurch, dass die Nutzendenbewertungen eine so zentrale Stelle im Geschäftsmodell des Unternehmens haben und die verzerrten Bewertungen durch die Nutzenden in die Unternehmensentscheidungen einfließen, sind mittelbare Diskriminierungen wahrscheinlich. Die Autor*innen

gehen davon aus, dass derartige Diskriminierungsrisiken für alle Plattformen relevant sind, bei denen ein System der Bewertungen durch Nutzende eine verteilte Belegschaft reguliert.

4.8 Staatliche Sozialleistungen und Aufsicht

Beispiel 27: Fortsetzung von Ungleichbehandlungen in Vorhersagesystemen der staatlichen Aufsicht

Altenburger und Ho (2018) zeigen zunächst, das Bewertungen von asiatischen Restaurants auf der Onlineplattform Yelp verzerrt waren, indem asiatische Restaurants unverhältnismäßig schlechter durch die Restaurantbesuchenden bzw. Kund*innen bewertet wurden als andere. Dazu haben sie für die Regionen New York und King County in den USA die Bewertung auf der Plattform bzw. Beschwerden bei Notrufdiensten durch Kund*innen und Daten zu Inspektionen durch Gesundheitsbehörden verglichen. Die Autor*innen zeigen sodann, dass sich derartige Ungleichbehandlungen durch verzerrte Bewertungsergebnisse von den privaten in den öffentlichen Bereich fortsetzen können, dadurch, dass die Bewertungen durch Kund*innen auch für öffentliche Aufgaben der Lebensmittel- und Gesundheitskontrolle verwendet werden. Nach Meinung der Autor*innen ist dies ein allgemeiner Entwicklungstrend bei der staatlichen Regulierung, das staatliche Kontrollen durch Big-Data-Analysen auf Basis von Bewertungsdaten aus Plattformen ersetzt werden. Die Fortsetzung von Verzerrungen geschieht dann, wenn algorithmenbasierte Prognosesysteme („predictive analytics“) für die staatliche Aufsicht eingesetzt werden, die auf der Datenbasis der verzerrten Kund*innenbewertungen entwickelt und betrieben werden. Zur Demonstration haben sie die Konsequenzen der Verwendung eines maschinellen Lernverfahrens zur Vorhersage auf Basis der verzerrten Bewertungen abgeschätzt. Erfolgt nach ihrer Meinung der Ersatz von staatlichen Lebensmittel- und Gesundheitskontrollen der Restaurants mit Vorhersagesystemen auf Basis der Auswertungen von verzerrten Kund*innenmeinungen werden nach ihrer Meinung verzerrt gebildete Algorithmen zu Regulierungsinstrumenten.

Beispiel 28: Ungleichbehandlung bei einem System zur Prävention von Kindesmisshandlungen

Das Allegheny Family Screening Tool (AFST) ist ein System zur Vorhersage und Bestimmung von präventiven Eingriffen bei potenziellen Fällen von

Kindesvernachlässigung und Kindesmissbrauch, das im Bezirk Allegheny County des US-Bundesstaates Pennsylvania eingesetzt wird (Einsatz seit 2016, Revision in 2018). Bei jedem Telefonanruf wird für jeden Fall ein Score von 1 bis 20 durch das Risikobewertungssystem vergeben, der den Bearbeitenden angezeigt wird. Dabei bedeutet ein Score von 20 das höchste Risiko. Die Anrufe bzw. Hinweise kommen in der Regel aus der Gemeinde, d. h. von Nachbar*innen oder Lehrkräften.

Für die Bestimmung der Risikoscores werden Ersatzinformationen bzw. Proxies verwendet (Eubanks 2017: 143–144). Das sind Daten zu Fällen in der Vergangenheit: (a) erneute Hinweise aus der „Gemeinde", d. h., wenn innerhalb von zwei Jahren ein erneuter Hinweis zum selben Kind eingegangen ist, zu dem beim ersten Anruf keine nähere Untersuchung vor Ort gestartet wurde („screened out") oder (b) Herausnahme aus der Familie („child placement"), wenn nach einem Anruf eine nähere Untersuchung gestartet wurde, die dazu geführt hat, dass das Kind aus der Familie entfernt und einer Pflegeeinrichtung zugewiesen werden musste. In den algorithmischen Schlussfolgerungen spiegeln sich also „soziale" vergangene Bewertungen der Familien durch die Gemeinde, die Behörde und die Gerichte wider. Das Modell könne auch nur Vorhersagen über künftige Hinweise und künftige Herausnahmen aus Familien liefern, nicht aber über die künftigen tatsächlichen Misshandlungen von Kindern. Zudem fehlen wichtige Variablen in der Modellbildung, wie geografische Isolierung. Das System vergibt unverhältnismäßig hohe Risikoscores an Familien, die bereits oft Sozialleistungen in Anspruch genommen haben, und erzeugt Ergebnisse mit Ungleichbehandlung nach ethnischer Herkunft (Eubanks 2017: 143–144; Courtland 2018). Für die Vorhersagegenauigkeit wurde ein Qualitätsmaß für Klassifikationen ermittelt, dem Wert für ROC AUC[62] , der mit 76 Prozent angegeben wurde (Eubanks 2017: 145). Da es bei vorhergehenden Versionen des Systems Probleme gab, wurde es einer Revision unterzogen und eine Gruppe von Forschenden mit der Überprüfung beauftragt (Chouldechova u. a.

[62] Die so genannte ROC-Kurve (Receiver Operating Characteristic, ROC) wird in einem Diagramm gebildet, in dem man die Richtig-Positiv-Rate auf der Ordinaten und die Falsch-Positiv-Rate auf der Abszisse einträgt. Bei einem Vergleich der Qualität von Klassifikatoren wird die Fläche unter der Kurve (Area under the Curve, AUC) betrachtet. Ein perfekter Klassifikator hat einen Wert von 1, und ergäbe einen Punkt in der linken, oberen Ecke des Diagramms, und ein völlig zufälliger einen Wert von 0,5 bzw. würde die Diagonale in dem Diagramm darstellen Géron (2018: 92–94). Man ist bestrebt, einen möglichst hohen ROC AUC-Wert zu erhalten. Die ROC AUC-Wert veranschaulicht den Kompromiss zwischen den Treffern bzw. richtig-positive Klassifizierungen und den „Kosten" bzw. falsch-positiven Klassifizierungen.

2018). Sie sehen in dem Gebrauch von „Predictive Analytics“ das Risiko von negativen Verstärkungseffekten, da von einigen regionalen Personengruppierungen („communities“), vor allem ärmere oder bestimmte ethnische Gruppierungen, mehr Daten in staatlichen Einrichtungen gespeichert sind, weil sie z. B. in Systemen der Sozialhilfe erfasst sind, und dadurch als hochriskant ausgewiesen werden und häufiger überprüft werden (Chouldechova u. a. 2018: 2).

Beispiel 29: Risiko der unmittelbaren und mittelbaren Diskriminierung bei einem System zur Einteilung von Arbeitslosen

In einer Studie der polnischen Nichtregierungsorganisation „Fundacja Panoptykon“ wurde auf Diskriminierungsrisiken durch ein System der Arbeitsvermittlung hingewiesen (Niklas, Sztandar-Sztanderska & Szymielewicz 2015). In 2014 wurde in Polen durch das Arbeitsministerium ein System zur Arbeitsvermittlung eingeführt, das Arbeitssuchende auf Basis der Bildung von Profilen und Scores in drei Kategorien („Profile I–III“) einteilt. Dadurch sollen die Ferne zum Arbeitsmarkt und die Bereitschaft, den Arbeitsmarkt zu betreten oder in diesen wieder einzutreten, bestimmt werden (Niklas, Sztandar-Sztanderska & Szymielewicz 2015: 11). Die Einteilung der jeweiligen Arbeitssuchenden in eine der drei Kategorien bestimmt die Art des anzuwendenden Arbeitsmarktprogramms (z. B. Vermittlung eines Arbeitsplatzes, Berufsausbildung, Lehre). Dabei werden 24 Merkmale, die Beschäftigungslose charakterisieren sollen, verarbeitet. Sie stammen aus der Registrierung der Beschäftigungssuchenden in den Arbeitsvermittlungszentren und den computergestützten Gesprächen mit den Sachbearbeiter*innen. Unter den erfassten und verarbeiteten Merkmalen sind Alter, Geschlecht und Grad der Behinderung (Niklas, Sztandar-Sztanderska & Szymielewicz 2015: 5, 11) sowie Angaben über Zeiten der Kinderbetreuung und Pflege bedürftiger Personen (ebd., S. 21). Ferner werden auch Angaben zum Bedürfnis und der Bereitschaft, eine Arbeit aufzunehmen, z. B. das eigene Engagement, eine Beschäftigung zu finden, die Bereitschaft, den Bedürfnissen des Arbeitsmarktes zu entsprechen, Flexibilität oder die frühere oder aktuelle Bereitschaft, mit den relevanten Arbeitsmarktbehörden zu kooperieren, erfasst (ebd., S. 11).

In der Praxis der Arbeitsvermittlungszentren wurden die Systeme von den Sachbearbeiter*innen unterschiedlich gehandhabt, u. a. wurde das Computersystem als Träger der endgültigen Entscheidung angesehen, teils wurde das Profiling als Teil einer umfassenden Untersuchung betrachtet, teils wurde versucht, die Profile entsprechend der Erwartungen der arbeitslosen

Person anzupassen (Niklas 2018). Nach Angaben der Autor*innen ist eines der Probleme des Systems, dass rechtliche Vorgaben, z. B. zur Kenntnisnahme über die Arten und den Umfang der zu verarbeitenden Daten und zur Anpassung der Angaben und des Profils nicht korrekt in den Funktionen und Regeln des Systems und seines Betriebs übersetzt worden sind. So sind beispielsweise die Anpassung und Korrektur der Eingaben erschwert (Niklas, Sztandar-Sztanderska & Szymielewicz 2015: 17; Niklas 2018). Nach Meinung der Autor*innen kann die Anwendung sowohl zu einer unmittelbaren als auch mittelbaren Diskriminierung führen. Da die Kategorienzuordnung auch mit den geschützten Merkmalen Alter, Geschlecht und Behinderung gebildet werden, läge eine unmittelbare Diskriminierung vor. Die mittelbare Diskriminierung ergäbe sich aus der Anwendung der Merkmale „Zeiten für Kinderbetreuung und Pflege", die statistisch gesehen öfter Frauen betreffen (Niklas, Sztandar-Sztanderska & Szymielewicz 2015: 21). Konkret hätten dann Beschäftigungssuchende, die der ungünstigsten Kategorie „Profil III" zugeordnet werden, eine geringere Chance, eine Fördermaßnahme zu erhalten. Diese Einschätzung der Chancen beruht auf statistischen Auswertungen und Erfahrungen (Niklas, Sztandar-Sztanderska & Szymielewicz 2015: 25–27).

Ferner haben Beschäftigungssuchende keinen ausreichenden Zugang zu den Daten und Auswertungen, die ihre Beurteilung und Zuordnung zu einer der drei Kategorien ausmacht. Dadurch können sie ihr Recht auf Schadenersatz beim Vorliegen von Diskriminierung nur schwierig geltend machen. Denn selbst bei Umkehrung der Beweislast auf die der Diskriminierung Beschuldigten, müssen die von einer möglichen Diskriminierung betroffenen Personen beweisen, dass eine Wahrscheinlichkeit der Diskriminierung besteht (Niklas, Sztandar-Sztanderska & Szymielewicz 2015: 21). Des Weiteren liegt nach Meinung der Autor*innen eine Verletzung des Rechts auf Verfahrensgerechtigkeit vor, da keine klaren Verfahren für den Widerspruch, die Einbringung der Meinung der betroffenen Personen sowie der Forderung nach Überprüfung vorgesehen seien (Niklas, Sztandar-Sztanderska & Szymielewicz 2015: 22f.). Nach Niklas ist jedoch die Abschaffung des Systems geplant (Niklas 2019).

Beispiel 30: Diskriminierungsrisiko bei Einteilung von Arbeitslosen

Medienberichten zufolge teilt ein in österreichischen Zentren der Arbeitslosenvermittlung (Arbeitsmarktservice, AMS) im Testbetrieb[63] befindliches System Arbeitssuchende in drei Kategorien der Integrationschancen auf dem Arbeitsmarkt ein (hohe, mittlere, niedrige Chance), um auf dieser Basis die Zuteilung von Mitteln für Qualifizierungsmaßnahmen, wie z. B. Schulungen oder Zusatzausbildungen, zu empfehlen (Fanta 2018). Die Integrationschancen werden als Wahrscheinlichkeit der Arbeitsaufnahme errechnet. Nach Angaben des Unternehmens Synthesis Forschung GmbH, das das System entwickelt hat, liegt der Berechnung der Wahrscheinlichkeiten ein logistisches Regressionsmodell zugrunde (Holl, Kernbeiß & Wagner-Pinter 2018). Die Trefferquote liegt nach Angaben des Vorstands der AMS bei 85 Prozent, was bedeutet, dass ca. 50.000 Menschen pro Jahr falsch eingeordnet werden. Bei diesem System liegt kein vollautomatisiertes Entscheidungssystem vor, denn das System gibt nur Empfehlungen ab (Wimmer 2018b, 2018a).

Kritisiert wurde zum einen, dass die Sachbearbeiter*innen der AMS tendenziell die Empfehlungen des Systems übernehmen, da insbesondere zusätzliche Evaluationen erfolgen müssen, wenn die menschliche Bewertung der Sachbearbeiter*innen von der Computerempfehlung abweicht und ein Herunterstufen der betroffenen Person vorgenommen werden soll (Wimmer 2018b). Zum anderen setzt die Kritik daran an, dass sich für bestimmte Personengruppen eine Benachteiligung ergeben könnte, denn der Algorithmus berücksichtigt nicht nur die geschützten Merkmale Geschlecht, Alter und Staatsangehörigkeit, sondern es werden auch Gewichtungen verwendet, die z. B. Frauen eine niedrigere Punktzahl zuweisen als Männern. Weitere, kontrovers diskutierte Merkmale sind Betreuungspflichten, also etwa Zeit für Kinderbetreuung, oder gesundheitliche Beeinträchtigungen (Szigetvari 2018).

Allerdings wurde von Seiten der AMS der Kritik entgegengehalten, dass die Einstufung in eine Kategorie der Arbeitsmarktchancen nicht bedeute, dass betroffene Gruppierungen bei den Qualifizierungsmaßnahmen schlechter gestellt seien. Hier sind vor allem Strategien und Programme der Förderungen von benachteiligten Gruppierungen zu berücksichtigen, die in den Entscheidungen über Qualifizierungsmaßnahmen einfließen (Wimmer 2018a; Salzburger Nachrichten 2019). Auch inwieweit die Kritik greift, dass die Bildung

[63] Stand Januar 2019.

des Algorithmus auf der statistischen Auswertung von vergangenheitsbezogenen Daten beruhe, wodurch es zu einer Verstetigung von Ungleichheiten auf dem Arbeitsmarkt kommen könnte, wird sich vermutlich erst durch die akkumulierten Förderentscheidungen im Hinblick auf die Umsetzung von Förderprogrammen beurteilen lassen, die angeblich gerade gegen eine weitere Benachteiligung ohnehin benachteiligter Gruppierungen gerichtet sind.[64]

4.9 Bildungswesen

Beispiel 31: Diskriminierungsrisiken bei einem System der Studienplatzvergabe

Nach Angaben der französischen Antidiskriminierungsstelle und den Dokumentationen des öffentlichen Verfahrens zufolge müssen beim französischen System der Studienplatzvergabe „Parcoursup" Studieninteressierte in einem Onlinesystem zehn bis 20 Wünsche für Studienorte und Angaben zu ihrer Person eingeben. Das System auf nationaler Ebene vermittelt zwischen den Wünschen und den jeweiligen Aufnahmekapazitäten der Hochschuleinrichtungen (öffentliche und private). Auf lokaler Ebene hat jede Hochschuleinrichtung ein eigenes System, das die Anfragen verwaltet und Studierende aussucht. Kritik besteht zum einen an der Intransparenz der Entscheidungsverfahren. Die Regierung machte auf nationaler Ebene zwar den Quellcode des Matching-Algorithmus öffentlich, die Sortieralgorithmen auf lokaler Ebene der jeweiligen Hochschuleinrichtungen waren jedoch nicht öffentlich. Zu den persönlichen Daten, die bei der Anmeldung angegeben werden müssen, gehören einerseits das Einkommen und der Wohnort, sodass Befürchtungen bei den Klägern über Diskriminierungen gegen weniger wohlhabende Studieninteressierte oder solchen aus Vororten bestanden.[65]

Ein Zusammenschluss von gewählten Vertretenden, Studierendenvertretenden, Lehrkräften, Jurist*innen brachte den Fall gegenüber der französ-

[64] So fordert die Volksanwaltschaft eine Überprüfung und parlamentarische Diskussion des AMS-Systems. Vgl. DerStandard (2019). Kritisch sehen dies jedoch Fröhlich & Spiecker genannt Döhmann (2018), die bereits in der Verwendung einer niedrigeren Gewichtung bei Frauen eine Diskriminierung sehen, ebenso Allhutter (2019), die eher von sich selbst verstärkenden Prozessen ausgeht, vor allem für Personen, auf denen mehrere Merkmale zutreffen, denen geringere Arbeitsmarktchancen zugeschrieben werden.

[65] Angaben durch Mitarbeiter*in von Défenseur des Droits, per E-Mail, November 2018 und März 2019.

sischen Antidiskriminierungsstelle Défenseur des Droits (DDD) vor, die eine rechtliche Untersuchung durchführte. Die Untersuchung führte zu zwei Entscheidungen durch DDD in Form von Empfehlungen an das Bildungsministerium: Entscheidung Nr. 2018–323 bezieht sich auf die besondere Berücksichtigung von Studierenden mit Behinderungen.[66]

Die zweite Entscheidung Nr. 2019–021 aus dem Januar 2019 bezieht sich auf die „lokalen Algorithmen" und empfiehlt unter anderem alle Informationen bezüglich der Verarbeitung, einschließlich der zu den Algorithmen, und der Bewertung von Bewerbungsunterlagen durch die lokalen Kommissionen der Hochschuleinrichtungen vorab zu veröffentlichen. Dies soll die Transparenz des Verfahrens gewährleisten und es den Bewerbenden erlauben, ihre Entscheidungen in voller Kenntnis der Fakten zu treffen. Mit der Entscheidung wird zudem daran erinnert, dass die Verwendung des Merkmals der Herkunftsschule bei der Auswahl der Bewerbenden durch Bevorzugung bestimmter Bewerbender oder Ausschluss anderer nach dem geografischen Standort der Schule als diskriminierende Praxis angesehen werden kann, wenn sie zum Ausschluss von Bewerbenden auf dieser Grundlage führt.[67]

Während der Untersuchung erhielt DDD keinen Zugriff auf die Algorithmen der „lokalen Ebene", aber allgemeine Informationen über deren Nutzung. Grundsätzlich hat DDD die rechtliche Möglichkeit, alle Informationen zu sammeln, die notwendig für die Fakten erscheinen, die ihnen gezeigt werden. Einschränkungen der Informationen aufgrund des Geheimnisschutzes, die nur bei Geheimhaltung zur nationalen Sicherheit, Staatssicherheiten oder Außenpolitik bestehen, waren in diesem Fall nicht relevant. Die Systeme auf „lokaler Ebene" bei den einzelnen Hochschuleinrichtungen sind freiwillig und ohne [übergreifendes] System eingeführt worden und haben die Aufgabe, die Bewerbungen in eine Vorab-Rangfolge zu bringen, bevor die eigentliche Auswahl durch Auswahlkommissionen erfolgt. Es handelt sich dabei um Unterstützung von Entscheidungen. DDD konnte während ihrer Untersuchung kein System mit vollständig automatisierter Datenverarbeitung identifizieren. Die Antidiskriminierungsstelle DDD arbeitete mit der französischen Datenschutzbehörde CNIL (Commis-

---

66 Angaben durch Mitarbeiter*in von Défenseur des Droits, per E-Mail, März 2019. Interpretierende Ergänzung durch den Autor in eckigen Klammern.

67 Siehe Décision 2019-021 du 18 janvier 2019 relative au fonctionnement de la plateforme nationale de préinscription en première année de l'enseignement supérieur (Parcoursup), abrufbar unter https://juridique.defenseurdesdroits.fr/index.php?lvl=notice_display&id=27285 (letzter Abruf am 26.3.2019).

sion Nationale de l'Informatique et des Libertés) zusammen. Aus der Rechtslage für derartige Zulassungsentscheidungen an Hochschulen schlussfolgerten sie, dass Prüfungskommissionen für Antworten an die Bewerbenden eingesetzt werden müssen und daher solche Entscheidungen nicht vollständig automatisierte sein können. Das „Parcoursup"-System auf „nationaler Ebene" gehört zu den Algorithmen, die durch ministeriellen Beschluss autorisiert werden müssen, nach vorhergehender Stellungnahme durch die Datenschutzbehörde CNIL mit Aussprache und Veröffentlichung. Nach Prüfung der Einhaltung der rechtlichen Vorgaben zur Transparenz, dem Recht auf menschliche Intervention, der gesammelten Datenarten und des Personenkreises mit ermächtigtem Datenzugriff gab CNIL eine Stellungnahme ab, die die Einführung des „Parcoursup"-Systems befürwortete. Allerdings wurden die Systeme auf lokaler Ebene nicht an die Überprüfung und Stellungnahme durch CNIL gesandt.[68]

4.10 Polizeiwesen

Beispiel 32: Verstärkung der Ungleichbehandlung bei einem System der vorausschauenden Polizeiarbeit

In einer Simulationsstudie untersuchten Lum und Isaac das System „PredPol" zur vorausschauenden Polizeiarbeit („predictive policing"), in dem Gebiete mit höherer Verbrechenswahrscheinlichkeit vorhergesagt werden sollen (raumbezogene Vorhersagen). Dabei gingen sie von dem Effekt aus, dass Verzerrungen in den (Trainings-)Datensätzen zu verzerrten Vorhersagen von Verbrechen führen und entsprechend zu mehr Einsätzen in den Gebieten, die in den Polizeistatistiken bereits überrepräsentiert sind. Durch die dann wahrscheinlich vermehrten Einsätze in denselben Gebieten werden zusätzliche Straftaten beobachtet, wodurch es zu einer Bestätigung der vorhergehenden Annahmen über die Verteilung der Straftaten kommt. In der Simulationsstudie sollte gezeigt werden, wie stark diese Verzerrung ist. Lum und Isaac betonen, dass Datensätze der Polizei keine vollständigen Erhebungen aller strafbaren Handlungen sind und sie auch keine Zufallsstichprobe repräsentieren. Um festzustellen, wie stark die Verzerrungen in den Datensätzen der Polizei waren, haben sie die Aufzeichnungen der Polizei mit einer vollständigen Kriminalitätsstatistik in Beziehung gesetzt, die aus dem „National Survey on Drug Use and Health" stammte. Diese Daten wur-

[68] Angaben durch Mitarbeiter*in von Défenseur des Droits, per E-Mail, März 2019. Interpretierende Ergänzungen in eckigen Klammern durch den Autor.

den auf die regionale Bevölkerungszusammensetzung der Stadt Oakland simuliert, wodurch Schätzungen über Drogenstraftaten auf regional hochaufgelöster Ebene für einzelne Gebiete der Stadt möglich waren, die nach ihrer Meinung ein genaueres Bild des Drogenmissbrauchs lieferten als die Polizeidaten zu Verhaftungen (Lum & Isaac 2016).

Das Ergebnis der Simulation war, dass die Drogenstraftaten viel verteilter über die Stadt sein müssten, die tatsächlichen Festnahmen wegen Drogenstraftaten der Polizei sich aber auf ein eng begrenztes Gebiet konzentrierten. Für die Untersuchung des Systems zur vorausschauenden Polizeiarbeit der Firma PredPol nutzten sie eine öffentlich zugängliche Version des Algorithmus. Nach Firmenangaben werden zur Vorhersage nur drei Datenpunkte verwendet: der vorhergehende Typ des Verbrechens, der Verbrechensort und die Zeit des Verbrechens. Der Algorithmus wurde auf die Polizeistatistiken eines Jahres angewandt und zeigte die Gebiete mit bereits hohen, überrepräsentativen Polizeieinsätzen („over-policed") als Gebiete mit jeweils hohen Vorhersagewerten an. Im Vergleich zu der simulierten Verbrechensverteilung, bei der eine Gebietsverteilung mit einem höheren Anteil von Personen mit Weißer Hautfarbe vorlag, waren bei den durch das PredPol System gelieferten Vorhersagen weniger Gebiete betroffen und vor allem die, in denen ein höherer Anteil von Schwarzen Personen lebten. Zudem simulierten sie die Situation, wenn Anreize für weitere Polizeieinsätze in „vorhergesagte" Gebiete bestünden, mit dem Ergebnis, der Bestätigung des oben genannten Effekts (ebd.).

Die Studie wurde kritisiert, da das PredPol System nicht zur Bekämpfung von Drogenverbrechen entwickelt wurde und die Verwendung von Drogendaten daher inkorrekt gewesen sei, sowie dass das System nicht in Oakland angewendet werde (Ferguson 2017). Die Autor*innen entgegneten, dass es um den Nachweis der Möglichkeit des Verstärkungseffekts ging, bei dem der problematische Punkt ist, dass Daten zu Verbrechen verwendet werden, die durch die Polizei (quasi als Nebentätigkeit) erzeugt werden, und dass diese keine Repräsentation aller Verbrechen sind, dass die Polizei nicht bei allen Verbrechen benachrichtigt wird und dass die Polizei nicht alle Verbrechen, auf die sie reagiert, dokumentiert. „Police-recorded crime data is a combination of policing strategy, police-community relations, and criminality." (Isaac & Lum 2018: ohne Seitenangabe). Zudem sei das System mangels Nachweis von Verbesserungen in der Verbrechensbekämpfung nach Erprobung abgeschafft worden (ebd.).

Beispiel 33: Diskriminierung nach ethnischer Herkunft bei einem System zur vorausschauenden Polizeiarbeit

Brantingham u.a. (2018) berichten von einer empirische Untersuchung (randomisierte kontrollierte Studie) zur Aufdeckung von möglichen Diskriminierungen nach ethnischer Herkunft (bzw. Minoritäten), die Teil einer umfassenden Begleitforschung zur Einführung eines Predicitive Policing Systems in Los Angeles war. Dazu wurden die Auswirkungen von Einsatzplänen der Polizeikräfte mit Kriminalitätsprognosen für jeweilige Einsatzgebiete, die von dem algorithmischen System erstellt wurden, mit denen eines menschlichen Planenden verglichen. Den Polizeikräften wurde vermittelt, dass die Einsatzgebiete die Gebiete mit den höchsten Kriminalitätsraten für ihre Schicht sind. In der umfassenden Begleitforschung konnte die Reduktion der Kriminalitätsrate sowohl bei den algorithmischen als auch bei den menschlichen Prognosen nachgewiesen werden, wobei der Rückgang bei den algorithmischen Prognosen höher ausfiel. Die Ergebnisse der Untersuchung zu Diskriminierungsrisiken zeigten, dass die Nutzung des algorithmischen Systemes keine Unterschiede bei den Festnahmen bei Personen von Minoritäten bewirkte. Dagegen habe sich die Rate der Festnahmen in den Gebieten, in denen das algorithmische System eingesetzt wurde, erhöht (über alle Bevölkerungsgruppen hinweg).

Beispiel 34: Datensätze mit diskriminierenden Polizeipraktiken bei Systemen der vorausschauenden Polizeiarbeit

Im Zusammenhang mit Systemen der vorausschauenden Polizeiarbeit („predictive policing“) erweitern Richardson u.a. den Begriff „dirty data“, der neben fehlenden oder falschen Daten oder nicht-standardmäßiger Repräsentation der Daten nun auch „schmutzige“ Datensätze enthalten, die auf korrupten, diskriminierungsgeneigten oder unrechtmäßigen Polizeipraktiken beruhen. Zudem können sie auch durch bewusste Manipulationen erzeugt und verändert worden sein. Besonders problematisch ist es, wenn die Systeme der vorausschauenden Polizeiarbeit in der Entwicklung und Anwendung auf derartigen „schmutzigen“ Datensätzen der Polizei beruhen. Auch diese Autor*innen betonen, dass Polizeidaten keine objektiven Daten seien oder tatsächliches kriminelles Verhalten oder Muster abbilden, sondern lediglich die Praktiken, Grundsätze und Programme, Voreingenommenheit sowie die politischen oder finanziellen Notwendigkeiten einer bestimmten Abteilung widerspiegeln (Richardson, Schultz & Crawford 2019: 8).

Ihre Studie beruht auf den Dokumenten und Ergebnissen von rechtlichen Untersuchungen, die durch die Regierung beauftragt wurden, oder auf gerichtlichen Vergleichen unter der Beaufsichtigung des Bundesgerichtshofs, Konsensvereinbarungen oder anderen Vereinbarungen, die auf rechtlichen Untersuchungen basieren. Sie betrachteten dazu 13 regionale Verwaltungseinheiten bzw. Jurisdiktionen („jurisdictions“)[69], die Systeme der vorausschauenden Polizeiarbeit in den Untersuchungszeiträumen einsetzten oder zuvor in Erprobungen eingesetzt hatten. Die Besonderheit dabei ist, dass die Verwaltungseinheiten in den Zeiträumen unter rechtlicher Untersuchung standen oder Verfahren der gerichtlichen Einigung liefen, mit denen herausgefunden wurde, dass deren Polizeibehörden Korruptionen, Ungleichbehandlungen nach ethnischer Herkunft („racially biased“) oder andere illegale Praktiken ausübten. Bei neun davon wurden von den Autor*innen Beweise zusammengetragen, die belegen, dass die Systeme zu Zeiten der illegalen Polizeipraktiken eingesetzt wurden. Drei Verwaltungseinheiten (Chicago, New Orleans und Maricopa) wurden in detaillierten Fallstudien dargestellt, die zu den Schlussfolgerungen über die Problematik der Verwendung von schmutzigen Daten führten (Richardson, Schultz & Crawford 2019).

Die Autor*innen führen diese Situationen auf fehlende Aufsicht und Überprüfungsmaßnahmen bei der Sammlung, Analyse und Verwendung von Polizeidaten zurück, die weder durch eine Behörde noch durch die Herstellenden solcher Systeme erfolgte (ebd., S. 20). Ferner konnten in den Fällen keine Hinweise gefunden werden, dass Herstellende bzw. Anbietende der Systeme die genutzten Polizeidaten unabhängig überprüften (ebd., S. 7). Anbietende dieser Systeme, die selbst auf verzerrte Datensätze der Polizei hinweisen, würden nicht ausreichend die strukturellen und systematischen Fehler dieser Daten einbeziehen. Die Identifizierung und Korrektur solcher Fehler sei eine zu große, wenn nicht gar unüberwindliche Herausforderung und lässt bezweifeln, dass zwischen problematischen und weniger problematischen Datenkategorien unterschieden werden kann. Selbst wenn eine Unterscheidung möglich wäre, sei dies zudem jeweils nur für eine jeweilige Verwaltungseinheit möglich und lasse kaum vergleichende oder aggregierte Schlussfolgerungen zu (ebd., S. 8f.). Zudem ist ohne eine autorisierte und unabhängige Behörde („empowered and independent authority“, ebd., S. 24) zu erwarten, dass potenziell unrechtmäßige und diskri-

[69] Genannt werden Boston, Chicago, Ferguson, Miami, Maricopa County, Milwaukee, New Orleans, New York, Newark, Philadelphia, Seattle und Suffolk County.

minierende Polizeipraktiken sowie darauf aufbauende Daten unbehandelt und unkorrigiert bleiben können, vor allem da es nur wenige politische und institutionelle Anreize zur Selbstüberprüfung und Reform gäbe (ebd., S. 24f.).

4.11 Strafvollzug

Beispiel 35: Diskriminierung nach ethnischer Herkunft bei Systemen zur Ermittlung der Strafrückfälligkeit

Einer der am häufigsten zitierten Beispielfälle von algorithmischen Systemen zur Entscheidungsunterstützung ist das COMPAS System (Correctional Offender Management Profiling for Alternative Sanctions) des Unternehmens Northpointe (heute „equivant"). Es wird für die Risikoprognose zur Unterstützung von Richter*innen in vielen US-Bundesstaaten verwendet, um Risiken der Strafrückfälligkeit bei vorzeitigen Entlassungen zu bestimmen. Das COMPAS System nutzt vielfältige persönliche Merkmale und einen proprietären Algorithmus, der die Wahrscheinlichkeit einer erneuten Verhaftung für jede prozessual angeklagte Person bestimmt. Auf öffentlich verfügbaren Dokumentationen und Kriminalitätsstatistiken basierend, fanden Recherchen des Journalistenverbundes ProPublica (Angwin u. a. 2016) heraus, dass die Vorhersagen für Schwarze Angeklagte systematisch das Risiko überbewerteten. Von denen, die nicht wieder inhaftiert wurden, waren 45 Prozent der Schwarzen Angeklagten mit einem hohen Risiko gekennzeichnet worden. Im Vergleich dazu waren nur 23 Prozent der Weißen Angeklagten, die nicht wieder inhaftiert wurden, mit einem hohen Risiko versehen worden. Sie schlossen für die Genauigkeit der Vorhersagen, dass die Wahrscheinlichkeit für Schwarze fälschlicherweise als hohes Risiko gekennzeichnet zu werden, doppelt so hoch ist als für Weiße Angeklagte. In einer Reaktion des Unternehmens wurde das statistische Vorgehen von ProPublica kritisiert und die eigenen Berechnungsmethoden dargestellt. Sie zeigten, dass Personen mit ähnlichem Risikoscore, unabhängig ob Schwarze oder Weiße, dieselbe Wahrscheinlichkeit hatten, wieder inhaftiert zu werden (Dieterich, Mendoza & Brennan 2016). Es zeigte sich, dass beide Parteien jedoch unterschiedliche Vorgehensweisen und Fairness-Konzepte verwendeten, deren gleichzeitige Anwendung und Erfüllung nicht möglich war (Chouldechova 2017; Eckhouse u. a. 2019).

2016 kam es zu einem Gerichtsentscheid des Wisconsin Supreme Court, bei dem eine „due process violation" bei der Verwendung des COMPAS Systems

zur Risikobeurteilung festgestellt wurde.[70] Entscheidend für den ablehnenden Gerichtsentscheid war die Aushöhlung des Rechts auf ein Gerichtsurteil basierend auf genauen Informationen (Freeman 2016; Citron 2016). Dabei bestätigte zwar das Gericht, dass die Bewertung von Individuen auf Basis von Gruppendaten und generalisierenden statistischen Auswertungen grundsätzlich rechtlich problematisch ist, doch der Supreme Court habe dieses Problem damit relativiert, dass die Risikoscores des COMPAS Systems nur eine von mehreren Informationsgrundlagen der Richter*innen seien. An der Entscheidung des Supreme Courts wird kritisiert, dass sie damit nicht ausreichend die (möglicherweise dominierende) Bedeutung von Computerempfehlungen auf menschliche Entscheidungen bzw. dem Risiko des „automation bias" Rechnung trage (Citron 2016; Freeman 2016: 96). Zuvor wurde auch dargelegt, dass derartige Systeme nicht kontinuierlich auf ihre Genauigkeit überprüft (Klingele 2015) und nicht auf verborgene Verzerrungen untersucht worden waren (Starr 2014), was auch in der Entscheidung des Supreme Courts kritisch diskutiert wurde (nach Citron 2016).

Beispiel 36: Vergleich Risikoanalysesystem mit maschinellen Lernverfahren

Tolan u. a. (2019) vergleichen das System SAVRY („structured assessment of violence risk in youth"), einem Risikoanalysesystem zur Vorhersage der Rückfallwahrscheinlichkeit im Jugendstrafvollzug, mit Verfahren des maschinellen Lernens hinsichtlich Gruppenfairness bei den geschützten Merkmalen Geschlecht und Nationalität. Das SAVRY System berechnet einen Gesamtscore durch die Eingabe von Einschätzungen bewertender Expert*innen zu Risiko- und Präventivfaktoren der jeweiligen Betroffenen. Die Zuordnung zu Risikoklassen bei der letztendlichen Entscheidung wurde ebenso durch die Expert*innen vorgenommen, sie war demnach nicht algorithmisch. Die Expert*innen wurden zuvor darüber informiert, dass es üblicherweise unterschiedliche Rückfallraten für männliche oder weibliche Straftäter*innen gibt.

Das SAVRY System wurde mit Verfahren des überwachten maschinellen Lernens verglichen. Dabei erfolgte der Vergleich auf Basis von zahlreichen Inputinformationen zu demografischen Angaben und der Strafhistorie der Angeklagten und es wurde eine Datensatz zu Jugendstraftaten in Katalonien verwendet. Als Ergebnis wurde gezeigt, dass das SAVRY System als

[70] Siehe Urteil des Supreme Court of Wisconsin „State v. Loomis", (881 N.W.2d 749, 763–64 (Wis. 2016)).

„fair“ hinsichtlich verschiedener Fairnessmaße[71] betrachtet werden kann, während die Verfahren des maschinellen Lernens dazu tendierten, männliche und ausländische Angeklagte sowie solche bestimmter Nationalitäten zu diskriminieren.

Sie nutzten neben Datenanalysen auch Werkzeuge zur Interpretation von ML-Verfahren. Die Forschenden betonen das Spannungsverhältnis zwischen Vorhersagegenauigkeit, bei der die ML-Verfahren besser abschneiden, und Erfüllung von Fairnessmaßen, bei denen sie schlechter sind. Eine Erklärung sehen sie darin, dass die grundlegende Verteilung von Rückfalltäter*innen in den verschiedenen Bevölkerungsgruppen („base rates“) sich auf die Vorhersagen der ML-Verfahren auswirkt. Unter anderem würden die ML-Verfahren die empirischen Korrelationen zwischen Bevölkerungsmerkmalen und Rückfallraten übernehmen.

Zusätzlich diskutieren sie mögliche technische Gegenmaßnahmen[72], die jedoch zu weiteren Verschlechterungen führen würden. (a) Das Weglassen von geschützten Merkmalen würde sich als nutzlos erweisen, denn viele andere Merkmale weisen Korrelationen zu ihnen auf. (b) Die Verwendung von verschiedenen Schwellenwerten (der Fairnessmaße) für verschiedene geschützte Merkmale kann zu falschen Klassifizierungen führen, mit der Folge, dass unerkannte Rückfalltäter*innen die öffentliche Sicherheit gefährden oder Jugendliche fälschlicherweise inhaftiert würden. (c) Die Anpassung des Modells bzw. des Klassifizierungsalgorithmus, z. B. indem eine Art Korrekturvariable eingefügt würde ohne das der zugrundeliegende Mechanismus verstanden würde, kann wiederum zu anderen Diskriminierungen, Stigmatisierungen oder Ungerechtigkeiten führen.

---

[71] Dabei wurden die Fairnessmaße der gleichen demografischen Parität („demographic parity“) und gleiches Fehlerverhältnis („error rate balance“) untersucht. Siehe auch Abschnitt 6.1.1 zu Fairnessmaße. Bei der Vorhersage von Rückfallwahrscheinlichkeiten bedeutet eine gleiche demografische Parität, dass jede Person mit einem geschützten Merkmal dieselbe Wahrscheinlichkeit hat, als Rückfalltäter klassifiziert zu werden, wie eine Person aus einer Referenzgruppe. Ein gleiches Fehlerverhältnis bedeutet, dass jede Person mit einem geschützten Merkmal die gleiche Wahrscheinlichkeit hat, fälschlicherweise als Rückfalltäter klassifiziert zu werden, wie eine Person aus einer Referenzgruppe. Dabei wurden die gleiche Falsch-Negativ-Rate und die gleiche Falsch-Positiv-Rate untersucht. Vgl. Tolan u.a. (2019).

[72] Siehe auch Abschnitt 6.1.1.

4.12 Übergreifende Beispielfälle der künstlichen Intelligenz

Beispiel 37: Ungleiche Genauigkeit bei Gesichtserkennungssystemen

Die Forschenden Klare u. a. (2012) zeigen in einer experimentellen Untersuchung, dass sechs untersuchte Gesichtserkennungssysteme, die u. a. von Strafverfolgungseinrichtungen in den USA verwendet werden, systematisch mit geringerer Genauigkeit Personenbilder mit den Kennzeichnungen „Frauen", „ethnische Herkunft" und für Personen zwischen 18 und 30 Jahren erkannten. Unter den Systemen waren drei kommerzielle Systeme, die bei allen Tests schlechtere Genauigkeit aufwiesen. Aus den gesamten Testergebnissen ziehen die Autoren u. a. die Schlussfolgerung, dass bei der Auswahl der Trainingsdaten alle Personengruppierungen, die später mit den Systemen erkannt werden sollen, ausreichend vertreten sein sollten.

Im Hinblick auf diese Ergebnisse betonen Garvie u. a. , dass die Fehlerraten bei der Genauigkeit in der Praxis dazu führen können, dass häufiger unschuldige Personen der schlechter erkannten Gruppierungen kontrolliert werden könnten. In diesem Zusammenhang verweisen die Autor*innen darauf, dass Gesichtserkennungssysteme kaum auf mögliche Ungleichbehandlungen nach ethnischer Herkunft getestet wurden (Klare u. a. 2012:53–56).

Beispiel 38: Übernehmen geschlechtsbezogener Stereotypen bei maschineller Textanalyse

Bolukbasi u. a. zeigten bei einem Verfahren der Textanalyse mit maschinellem Lernen der natürlichen Sprache („natural language processing"), dass in den Texten verborgene geschlechtsbezogene Stereotypen sich in den Ergebnissen wiederholen. Das betrachtete Textanalyseverfahren „word embedding" wandelt Textdaten in (Zahlen-)Vektoren um, die dadurch für die weitere maschinelle Bearbeitung zur Verfügung stehen. Dabei wird das Verfahren im Hinblick auf gleichzeitiges Vorkommen von Wörtern in bestimmten Textkorpora trainiert und es wird nach bestimmten Zusammenhangsmustern zwischen Wörtern gesucht. Dabei bildet die Geometrie zwischen den Vektoren den semantischen Zusammenhang zwischen den Wörtern ab, aus dem unter anderem die Stereotypen in Form von errechneten Wortzuordnungen ersichtlich werden (Bolukbasi u. a. 2016).

In der Untersuchung wurde das öffentlich zugängliche Verfahren „Word-2Vec" anhand des Google News Textkorpus untersucht, der drei Millionen

englische Wörter enthält. Dabei wurde erwartet, dass dieser eigentlich wenig geschlechtsbezogene Verzerrungen enthalten würde, da er hauptsächlich durch professionelle Journalist*innen erstellt wurde. Allerdings wurden in zweifacher Weise stereotypartige Ergebnisse des maschinellen Lernverfahrens nachgewiesen. (a) Bei zugeordneten Beschäftigungsbezeichnungen wurde gezeigt, dass zum Wort „she" insbesondere Wörter wie „homemaker" (deutsch: jemand der den Haushalt führt), „nurse" (Krankenschwester, Kindermädchen), „receptionist" (jemand der am Empfang arbeitet), „librarian" (Bibliothekar*in), „socialite" (Persönlichkeit des öffentlichen Lebens) u. a. in einen Zusammenhang gebracht wurden, während für das Wort „he" dies Wörter wie „maestro" (Maestro), „skipper" (Kapitän*in), „protege" (Schützling), „philosopher" (Philosoph), „captain" (Kapitän*in) u. a. waren. Die Einschätzung, ob die rechnerisch zugeordneten Begriffe eher weiblich-stereotyp, männlich-stereotyp oder neutral waren, haben Crowdworker[73] übernommen. (b) Bei der maschinellen Bildung von Analogiepaaren, in der Form „man is to king as woman is to queen", lieferte das Verfahren Wörter, die in 29 von 150 Analogiewörtern als Geschlechterstereotyp angesehen wurden, während 72 der 150 Wörter als passend zum Geschlecht beurteilt wurden. Auch diese Beurteilung der rechnerischen Zuordnungen wurde von beauftragten Crowdworkern vorgenommen. Gleichzeitig stellen sie ein Verfahren vor, das geschlechtsbezogene Verzerrungen vermindern soll. Dazu haben sie für bestimmte Wörter die Verknüpfung zum Geschlecht verändert, wodurch z. B. der Begriff „nurse" eine weibliche und männliche Zuordnung bekam (Bolukbasi u. a. 2016).

Maschinelle Textanalysen werden in verschiedensten Anwendungen eingesetzt, wie der automatisierten Analyse von z. B. Dokumenten, Lebensläufen oder der schriftlichen Kommunikation in sozialen Netzwerken sowie der automatisierten Rangfolgenbildung bei Suchmaschinenergebnissen, Produktempfehlungen oder maschinellen Übersetzungen. Werden die so erzeugten „embedding"-Algorithmen, die stereotype Wortbeziehungen übernommen haben, dort eingesetzt, kann es zu problematischen Ergebnissen kommen, in dem Sinne, dass überkommene Geschlechterrollen fortgesetzt werden.

[73] Sogenannte „Crowdworker" sind Personen, die meist kleine Dienstleistungen der Computerarbeit über Onlineplattformen im Internet anbieten und bearbeiten, wie z. B. Testen von Webseiten oder Software, Schreiben von Texten, Kategorisieren von Fotos, Programmieren von Softwareteilen oder Designarbeiten. Meistens sind sie nicht fest bei einer Firma angestellt. Die Tätigkeit der Crowdworker wird auch als „Crowdsourcing" bezeichnet, in Anlehnung an das Outsourcing.

Beispiel 39: Übernehmen von kulturellen Stereotypen bei maschineller Textanalyse

Forschende (Caliskan, Bryson & Narayanan 2017) zeigten, dass gängige Verfahren des maschinellen Lernens (hier des „word embedding“) alltägliche kulturelle Stereotypen auch von Textdateien „erlernen“ können. Die Verfahren des maschinellen Lernens wurden u.a. an einem Textkorpus mit normaler menschlicher Sprache aus dem Internet, dem „common crawl“-Korpus, trainiert. Die Forschenden zeigten, dass bei Anwendung der maschinellen Lernverfahren auf Texten die Ergebnisse ebenso stereotyp sind, wie dies zuvor von anderen Forschenden für menschliches Verhalten durch Assoziationstests nachgewiesen wurde. Das zeigten sie dadurch, dass maschinell ermittelte Assoziationen zwischen Wörtern ähnlich ausfielen, wie die Assoziationen bei Menschen. So zeigten sie u.a., dass die ML-Verfahren weibliche Namen mehr mit Wörtern für „Familie“ als für „Karriere“ assoziierten, im Vergleich zu männlichen Namen (ähnliche Ergebnisse z.B. für Wörter für „Mathematik“ oder „Wissenschaft“ mit männlichen Begriffen und „Künste“ mit weiblichen Begriffen). Ferner bestätigten sie vorhergehende Forschungsergebnisse, dass Europäisch-amerikanische Namen mehr mit Begriffen für „angenehm“ assoziiert wurden im Vergleich zu Afroamerikanische Namen.

Daraus schließen sie, dass Systeme der KI, die die Eigenschaften von Sprache erlernen und reproduzieren, die vergangenen, teils stereotypen kulturellen Vorstellungen übernehmen. Dies ist insbesondere problematisch, wenn diese Systeme in gegenwärtigen Anwendungen, wie der Online-Textübersetzung, verwendet werden oder ihnen Entscheidungen überlassen werden, wie das Prüfen von Lebensläufen, bei denen Systeme mit übernommenen kulturellen Stereotypen mit Vorurteilen behaftete Ergebnisse produzieren würden (Caliskan, Bryson & Narayanan 2017).

Beispiel 40: Übernehmen von geschlechts- und ethnienbezogenen Stereotypen bei maschineller Textanalyse

Ähnlich wie beim vorhergehenden Beispiel zeigten Garg u.a. (2018), dass das maschinelle Lernverfahren „word embedding“ auch dazu geeignet ist, in großen Textmengen und über historische Zeitabläufe die jeweils verbreiteten geschlechtsbezogenen Stereotypen und Einstellungen zu ethnischen Minoritäten und deren Veränderungen im Zeitverlauf quantitativ zu erfassen. Dies zeigten sie anhand von Texten des 20. und 21. Jahrhunderts in den USA, u.a. mit den Datensätzen des Google Books/Corpus of Historical American English und dem New York Times Annotated Corpus. Beispielsweise wurden für 1910 mit dem Wort „Frauen“ die Assoziationen „char-

ming“ (charmant, reizend, …), „placid“ (friedfertig, gelassen …), „delicate“ (zart, empfindlich …), „passionate“ (leidenschaftlich), „sweet“ (süß, angenehm, niedlich …), „dreamy“ (verträumt …), „indulgent“ (nachsichtig, nachgiebig …), „playful“ (verspielt …), „mellow“ (umgänglich, heiter …) oder „sentimental“ (sentimental, empfindsam) ermittelt. Für 1990 waren es „maternal“ (mütterlich), „morbid“ (krankhaft …), „artificial“ (künstlich …), „physical“ (körperlich), „caring“ (warmherzig, mitfühlend …), „emotional“ (emotional …), „protective“ (fürsorglich, schützend …), „attractive“ (attarktiv, verlockend …), „soft“ (weich, schlaff …) und „tidy“ (ordentlich). Die Autor*innen weisen darauf hin, dass die automatisiert erlernten Stereotypen dann problematisch werden, wenn sie in „sensitiven“ Produkten und Diensten, wie z B. der Rangfolgebildung bei Suchmaschinen, Produktempfehlungen oder automatisierten Übersetzungen eingesetzt werden (Garg u. a. 2018: E3635).

Beispiel 41: Ungleiche Genauigkeit nach Geschlecht und Dialekt bei automatischen Untertiteln des Videodienstes YouTube

In einer wissenschaftlichen Untersuchung zeigte Tatman, dass der Dienst der Plattform YouTube, um automatische Untertitel bei hochgeladenen Videos zu erzeugen („automatic caption“), unterschiedliche Genauigkeiten aufweist, mit deutlich geringerer Genauigkeit der Erkennung der Sprache von Frauen und für Videos mit Personen mit schottischem Dialekt. Der Dienst beruht auf einem maschinellen Lernverfahren. Als einen der möglichen Gründe vermutet die Autorin unzureichende Trainingsdaten (Tatman 2017: 57).

Beispiel 42: Ungleiche Genauigkeit nach ethnischer Herkunft bei Systemen zur Spracherkennung

Blodgett u. a. untersuchten vier verbreitete Spracherkennungssysteme bzw. Cloud-Dienste (langid.py, IBM Watson, Microsoft Azure und Twitters interner Identifizierungsdienst) anhand des Dialekts des Afroamerikanischen Englisch, das im sozialen Netzwerk Twitter verwendet wurde.[74] Dazu nutzten sie einen öffentlich verfügbaren Twitter-Korpus mit 59,2 Millionen Tweets. Als Ergebnis haben sie Unterschiede in der Genauigkeit der Spracherkennung festgestellt, mit einer höheren Genauigkeit für die Weißen Autor*innen zugeordneten Textnachrichten gegenüber Textnachrichten, die Afroamerikanischen

[74] In einer vorhergehenden Publikation von Blodgett, Green & O‘Connor (2016) wird die Datenerstellung geschildert. Dazu mussten sie die Twitterbeiträge bzw. Tweets verschiedenen ethnischen Bevölkerungsgruppen zuordnen, wozu sie den Standort der Zielregion der Twitternachricht mit den demografischen Angaben der offiziellen Bevölkerungsstatistik abglichen. Ebenso wurde die Untersuchung der drei Spracherkennungssysteme langid.py sowie zwei interne Systeme von Twitter mit schlechterer Erkennung von Afroamerikanischem Dialekt dargestellt.

Autor*innen zugeordnet waren. Die Unterschiede sind insbesondere für kurze Textnachrichten besonders hoch. Da Daten aus sozialen Netzwerken häufig für Stimmungs- und Meinungsanalysen zu Produkten oder politischen Personen verwendet werden, sehen sie ein gesellschaftliches Problem darin, dass die Meinungen von Afroamerikaner*innen weniger gut erfasst werden als von Weißen Teilnehmer*innen (Blodgett & O'Connor 2017: 3).

Systeme der Spracherkennung werden in persönlichen Assistenzsystemen (z. B. Siri, Alexa, Amazon Echo), Chatbots oder automatisierten Telefonsystemen eingesetzt. Bei unzureichenden Trainingsdatensätzen, in denen bestimmte Bevölkerungsgruppen nicht ausreichend vertreten sind, können deren Dialekte oder Akzente nicht ausreichend erlernt werden, sodass Angehörende dieser Bevölkerungsgruppen in Anwendungen schlechter erkannt oder verstanden werden.

Beispiel 43: Ungleiche Erkennungsraten nach Geschlecht bei maschineller Meinungs- und Stimmungsanalyse

In einer Untersuchung von maschinellen Lernverfahren zur Meinungs- und Stimmungserkennung („sentiment analysis") auf Basis von Texten wurden Ungleichheiten zwischen Geschlechtern nachgewiesen, wobei die Verfahren besser dazu geeignet waren, die Stimmungen von Frauen festzustellen als diejenigen von Männern. Die Datenbasis bildeten Bewertungen von Hotels und Restaurants der Reiseplattform TripAdvisor.com für den Raum Großbritannien. Die aufgedeckten Geschlechterunterschiede würden bedeuten, dass bei Meinungs- und Stimmungsanalysen die Meinungen von Frauen leicht überrepräsentiert sind, da ein höherer Anteil von männlichen Stimmungen nicht erfasst wird (Thelwall 2018). Der Autor schränkt jedoch ein, dass die Ergebnisse für diesen spezifischen Datensatz nicht verallgemeinerbar sind. Jedoch zeigte er, dass die Verzerrungen nicht im Datensatz vorlagen, sondern erst als Ergebnisse der scheinbar objektiven maschinellen Lernverfahren zustande kamen. Er plädiert daher bei Überprüfungen der Algorithmen nicht nur dafür, den Input in Form der Daten und Algorithmen, sondern auch den Output des Systems zu testen, einschließlich der unterschiedlichen Gruppierungen, z. B. nach Geschlecht, ethnischer Herkunft etc. (Thelwall 2018). Automatisierte Meinungs- und Stimmungsanalysen werden im Marketing zur Produkt- oder Dienstleistungsbewertung eingesetzt und können in nahezu Echtzeit Stimmungen einer großen Anzahl von Nutzenden, z. B. von sozialen Netzwerken, erfassen.

Beispiel 44: Ungleiche Erkennung bei 219 Systemen der Stimmungs- und Meinungsanalysen

Kiritchenko und Mohammad untersuchten zusammen mit kooperierenden Einrichtungen der Evaluierungsstudie SemEval über 219 Systeme der Verarbeitung natürlicher Sprache zur Stimmungs- und Meinungsanalyse („sentiment analysis"), von denen einige unterschiedliche Resultate je nach Geschlecht oder ethnischer Herkunft erzielten. Die Datenbasis war ein einheitlicher Textkorpus, der „equity evaluation corpus" (EEC), der aus 8.640 englischen Sätzen gebildet wurde. Die Tests der Systeme wurden durch ein Netzwerk kooperierender Einrichtungen durchgeführt. Für die Untersuchung wurde eine systematische Quantifizierung der Ungleichheiten nach Geschlecht und ethnischer Herkunft durchgeführt. Die Ergebnisse wurden in Form der systematisch höheren oder niedrigeren Scores für die Stimmungsintensität („sentiment intensity score") dargestellt (Kiritchenko & Mohammad 2018).

Beispiel 45: Ungleiche Erkennung nach Hautfarbe bei Systemen zur Personenerkennung

Wilson u. a. testeten die ungleiche Vorhersagequalität bei Objekterkennung durch maschinelle Lernverfahren bei der Erkennung von zu Fuß Gehenden mit unterschiedlicher Hautfarbe. Sie untersuchten, ob die ungleichen Vorhersageraten („predictive inequity") auf die Tageszeit (bzw. Lichtverhältnisse), den Grad der Verdeckung von Personen oder auf Gewichtungen in der Zielfunktion des maschinellen Lernverfahrens zurückzuführen waren. Als Ergebnis zeigten sie, dass Standardmodelle der Objekterkennung, die an Standarddatensätzen trainiert wurden, eine höhere Genauigkeit für Bilder mit Personen mit „heller" Hautfarbe (Fitzpatrick Hauttypen in niedrigen Kategorien) als für „dunklere" Hautfarbe haben. Diese Ungleichheit nahm sogar zu, als die Forschenden verdeckte Personen (als mögliche Quelle der Genauigkeitsunterschiede) aus den Bildern entfernten (Wilson u. a. 2019).

Beispiel 46: Ungleiche Erkennung nach Hautfarbe bei kommerziellen Systemen der Gesichtserkennung

Buolamwini und Gebru zeigten durch Experimente mit kommerziellen Gesichtserkennungssystemen von Microsoft, IBM und Face++, dass die Erkennung von Gesichtern bei Geschlechtern und bei Personen mit unterschiedlicher Hautfarbe (eingeteilt nach der Fitzpatrick Hauttypkategorisierung) unterschiedlich gut funktionierte. Die Systeme erkannten grundsätzlich männliche Gesichter (8,1 Prozent Fehlerrate) besser als weibliche (20,6 Prozent Fehlerrate), ebenso Gesichter mit „heller" Hautfarbe (11,8 Prozent Fehlerrate) besser als Personen mit „dunkler" Hautfarbe (19,2 Prozent Fehler-

rate). Für die Tests haben sie einen eigenen Vergleichsdatensatz mit Personenbildern von 1.270 afrikanischen und europäischen Parlamentarier*innen zusammengestellt, den „Pilot Parliaments Benchmark“ (PPB). Bestehende Vergleichsdatensätze für die Bestimmung der Genauigkeit der Systeme waren nach ihrer Meinung durch eine Überrepräsentation von männlichen und „hellhäutigen“ Personentypen und eine Unterrepräsentation von weiblichen und dunkelhäutigen Personentypen gekennzeichnet (Buolamwini und Gebru 2018).

Die Veröffentlichung des Papers von Buolamwini und Gebru hat zu Reaktionen von Firmen geführt. Beispielsweise hatte IBM den Datensatz für die Optimierung ihrer Software an den Pilot Parliaments Benchmark angepasst (Puri 2018). Auch andere Hersteller hatten ihre Genauigkeit der Gesichtserkennung verbessert. Diese Reaktionen wurden in einer weiteren Studie von Raji und Buolamwini analysiert und diskutiert, bei der die Systeme von IBM, Microsoft, Face++, Amazon und Kairos untersucht wurden (Raji & Buolamwini 2019).

Beispiel 47: Ungleiche Erkennung nach Hautfarbe bei einem kommerziellem System zur Gesichtserkennung

Die amerikanische Bürgerrechtsorganisation „American Civil Liberties Union“ (ACLU) fand bei einem Test des Gesichtserkennungssystems „rekognition“ des Unternehmens Amazon heraus, dass das System 28 Mitglieder des US-Kongresses falsch erkannt hat, d.h. fälschlich als Personen von Fahndungsbildern erkannt hat, und dass die Rate der Falscherkennung bei PoC (People of Color) überproportional hoch sei. Bei dem Test des öffentlich über ein Onlineangebot verfügbaren Systems wurden die Bilder mit einer Datenbank von 25.000 öffentlich verfügbaren Fotos von Personen bei Festnahmen abgeglichen. Das System wird nach Angaben der Bürgerrechtsorganisation durch Polizeikräfte in Oregon zum Abgleich von Aufnahmen aus sogenannten „body cams“ mit einer Datenbank von Fahndungsfotos eingesetzt. Zum Zeitpunkt der Berichterstattung lief eine Protestaktion gegen die Verwendung des Überwachsungssystems (Snow 2018).

5. Ursachen von Diskriminierungsrisiken

Im Folgenden werden zur Veranschaulichung **zwei Arten** von Diskriminierungsrisiken unterschieden, die jedoch miteinander verbunden sind:

(1) Diskriminierungsrisiken, die durch die Verwendung von Algorithmen aufgrund ihrer besonderen technischen Eigenschaften resultieren (siehe Abschnitt 5.1) sowie

(2) solche, die durch die Verwendung der algorithmen- und datenbasierten Differenzierungen und Entscheidungssysteme an sich entstehen und als gesellschaftliche Risiken auftreten (siehe Abschnitt 5.2).[75]

In diesem Zusammenhang weisen Powles und Nissenbaum darauf hin, dass die derzeitige Konzentration auf technische Lösungsansätze für durch Algorithmen und KI verursachte Probleme auch mit gesellschaftlichen Gefahren verbunden ist: (1) Gesellschaftliche Ungleichbehandlungen sind ein soziales Problem und Lösungsversuche mithilfe der technischen Logik der Automatisierung sind stets inadäquat. (2) Selbst bei erfolgreicher Lösung von technischen Problemen sagt dies noch nichts über die Legitimität des Verwendungszwecks aus. Würde beispielsweise das Problem, dass Personen mit bestimmter Hautfarbe bei einem Gesichtserkennungssystem schlechter erkannt werden, gelöst, so kann das Gesichtserkennungs- bzw. -identifizierungssystem immer noch für Überwachungsaktivitäten mit unverhältnismäßiger Einschränkung von Grundrechten des Persönlichkeitsschutzes genutzt werden. (3) Schließlich kann ein einseitiger Fokus auf technische Lösungen die Aufmerksamkeit und (Forschungs-)Ressourcen von der Lösung der eigentlichen gesellschaftlichen Ungleichbehandlungsprobleme abbringen (Powels & Nissenbaum 2018).

[75] Ähnliche Unterteilungen nehmen auch andere Autor*innen vor, vgl. Citron & Pasquale (2014), Zarsky (2016), Crawford u. a. (2016: 6-7), Britz (2008: 120–136), Gandy Jr. (2010), Eckhouse u. a. (2019).

5.1 Risiken bei der Verwendung von Algorithmen, Modellen und Datensätzen

Diskriminierungsrisiken können sich aus den besonderen technischen Eigenschaften der Verfahren der algorithmenbasierten Datenauswertungen und automatisierten Entscheidungenverfahren ergeben, wobei im Folgenden vor allem die des Data-Mining und des maschinellen Lernens im Vordergrund stehen (nach Calders & Žliobaitė 2013; Barocas & Selbst 2016; Kim 2016; Lehr & Ohm 2017; Zweig, Fischer & Lischka 2018; Schweighofer u. a. 2018; Zuiderveen Borgesius 2018; Favaretto, De Clercq & Elger 2019; Tolan 2018; FRA 2019).

5.1.1 Risiken bei der Entwicklung der Algorithmen und Modelle

Üblicherweise wird beim Data-Mining und maschinellen Lernen der Analysezweck bestimmt, d. h., was vorhergesagt oder geschätzt und wie es gemessen werden soll. Aus verbalen Beschreibungen des Analysezwecks muss eine berechenbare **Zielfunktion** bestimmt werden, die anhand von historischen oder anderen Daten optimiert werden soll. Dabei ist eine Zielfunktion ein mathematischer Ausdruck des Ziels des gesamten Verfahrens, meist in Form von Zielvariablen, die minimiert oder maximiert werden (Lehr & Ohm 2017: 671). Werden die Verfahren in Organisationen wie Unternehmen oder Behörden eingesetzt, so sind die Zwecksetzungen, Vorgaben und Anforderungen der Organisation in berechenbare Funktionen und Werte umzuwandeln. Vorgaben, wie z. B. die „Kreditwürdigkeit" von Antragstellenden, die „Leistungsfähigkeit" von Arbeitnehmenden oder den „Wert" von Kund*innen zu analysieren oder zu bestimmen, müssen in berechenbare Größen umgewandelt bzw. formalisiert werden. Systementwickelnde und Analysierende haben damit die Aufgabe, zu bestimmen, was z. B. „gute Kreditwürdigkeit" oder „gute Mitarbeitende" ausmacht (Barocas & Selbst 2016: 678f.).

Zur Bestimmung, was z. B. „gute" Arbeitnehmende ausmachen, stehen üblicherweise sehr viele und sehr unterschiedliche Zielvariablen zur Verfügung. Am Beispiel der Personalauswahl kann verdeutlicht werden, dass Diskriminierungsrisiken bereits bei der Auswahl der Zielvariablen entstehen können (Barocas & Selbst 2016: 680). Wenn beispielsweise Entscheidungssysteme der Personalauswahl auf Zielvariablen basieren, die mit der

Auswertung von Betriebszugehörigkeitszeiten gebildet werden, werden bestimmte Gruppierungen, die üblicherweise eine höhere Wechselrate bei Anstellungen aufweisen (wie z. B. Frauen), systematisch benachteiligt, auch wenn sie gleich gute oder bessere Leistungen erbringen können. Ebenso wird das Problem am Beispiel 24 (S. 53) eines Computersystems zur Zuordnung von Patient*innen zu Behandlungsprogrammen verdeutlicht, bei dem eine Zielfunktion, die einseitig auf Kosten ausgerichtet ist und Gesundheitsaspekte nicht ausreichend abbildet, zu Diskriminierungen nach ethnischer Herkunft führte. Mit anderen Worten kann man festhalten, dass bereits die Auswahl und Bestimmung der Zielvariablen zu einem Diskriminierungsrisiko werden können, das in Algorithmen und Computersysteme quasi „einprogrammiert" wird.

Auch durch Auswahl und Bestimmung der Kennzeichnungen der Kategorien („labels" oder „class labels") kann es zu Diskriminierungsrisiken kommen. Das ist für die spätere Phase der Anwendung (auf neue Datensätze) bedeutend, denn die automatisierte Zuordnung von Personen zu Kategorien oder Gruppen geschieht mit diesen Kennzeichnungen bzw. Labeln. Normalerweise werden bei Auswahl und Bestimmung der Kategorien bzw. Label alle möglichen Werte der Zielvariablen in sich gegenseitig ausschließende Kategorien unterteilt. Da die Auswahl der Zielfunktion mit Zielvariablen und die Kennzeichnungen der Kategorien bereits die Ergebnisse des maschinellen Lernverfahrens beeinflussen, bilden sie die Grundlage für Diskriminierungsrisiken in den anderen Schritten der Verfahren des Data-Minings bzw. maschinellen Lernens (Barocas & Selbst 2016: 678, 680). Die auf dieser Basis gebildeten Entscheidungsregeln reflektieren durch die subjektiven Entscheidungen tendenziell bestehende Vorurteile und Verzerrungen und führen systematisch zu Benachteiligungen (Barocas & Selbst 2016: 682): (a) Bei objektiven Kennzeichnungen ist die Wahrscheinlichkeit geringer, dass derartige Verzerrungen auftreten, denn zwischen verschiedenen Beteiligten besteht Einigkeit über das Erfüllen oder Nichterfüllen des Merkmals und individuelle Auslegungen sind nicht erforderlich (Calders & Žliobaitė 2013: 48). Beispielsweise können die Merkmale „Kredit zurückgezahlt oder nicht" oder „auf Alkohol getestet oder nicht" eindeutig bestimmt und von jedermann nachvollzogen werden. Insbesondere jedoch, wenn (b) die Auswahl der Kennzeichnungen subjektive Auslegungen enthalten, können Diskriminierungsrisiken auftreten. Bei diesen Merkmalen kann es nicht eindeutig und von jedermann einvernehmlich nachvollzogen oder geteilt sein, wie man das Merkmal definieren oder messbar machen kann, wie z. B. das Merkmal „Passfähigkeit" von Bewerbenden für eine Personalstelle (Calders & Žliobaitė 2013: 48).

Ferner kann es beim Trainieren von Modellen zu Risiken bei der **Auswahl der Einflussvariablen** kommen. Bei Verfahren des Data-Minings bzw. maschinellen Lernens werden Modelle erzeugt, die auch sehr viele Einflussvariablen (hier „features" genannt) enthalten können. Eine Auswahl („feature selection") ist dann notwendig, wenn es technisch unmöglich ist, alle Einflussvariablen einzubeziehen. Durch die Auswahlentscheidungen können aber Modelle erzeugt werden, die keinen ausreichenden Detaillierungsgrad mehr haben, um kritische Unterscheidungspunkte zu erkennen. Dadurch kann es zu systematisch höheren Fehlerquoten beim Erkennen von bestimmten Personengruppen kommen, die auch Träger von geschützten Merkmalen sein können. Die verwendeten Merkmale können zwar statistisch hinreichend, aber für die nicht ausreichend erkannte Personengruppe schlicht unzutreffend sein. Mit anderen Worten: Sie sind nicht allgemeingültig.[76] An dieser Stelle werden die Eigenheiten und Probleme der statistischen Diskriminierung besonders deutlich. Zugunsten von Effizienzzielen werden Ersatzvariablen zur Differenzierung herangezogen, die zu grob sind, um den Charakteristika betroffener Personen ausreichend gerecht zu werden. Es entsteht ein Generalisierungsunrecht (Weiteres in Abschnitt 5.2.1, ab S. 86).

5.1.2 Risiken bei der Zusammenstellung der Datensätze und Merkmale

In der Statistik und Informatik ist seit längerem bekannt, dass verzerrte Datensätze zu diskriminierenden Modellen führen. Ebenso treten sie bei Verfahren des Data-Minings und des maschinellen Lernens auf (Custers 2013; Barocas & Selbst 2016: 680-690; Lehr & Ohm 2017; FRA 2019). Unter- oder Überrepräsentation von Personengruppen sowie die Abbildung früherer Ungleichbehandlungen in den verwendeten Datensätzen sowie Korrelationen der Ersatzvariablen bzw. Proxies zu geschützten Merkmalen können zu Diskriminierungsrisiken bei den Ergebnissen der Differenzierungssysteme führen.

Diskriminierungsrisiken ergeben sich durch absichtliche oder unabsichtliche **Unter- oder Überrepräsentation von Personengruppen** oder das komplette Weglassen in ausgewerteten Datensätzen bzw. Trainingsdaten, mit

[76] Vgl. Barocas & Selbst (2016: 688) mit Verweis auf Schauer (2003).

anderen Worten, wenn zu bestimmten Personengruppen Daten nicht in einem Verhältnis vorliegen, das für eine korrekte Repräsentation notwendig wäre (Calders & Žliobaitė 2013: 47–49; Barocas & Selbst 2016: 684–686). Mögliche Ursachen dafür sind:

(1) Die Daten beziehen sich auf Situationen, in denen eine **Ungleichbehandlung und Ungleichverteilung** von Personengruppen bestand oder besteht. Wenn beispielsweise ein Modell durch die Auswertung von Daten zu bestehenden oder historischen Beschäftigungsverhältnissen gebildet würde, etwa um einen Zusammenhang zwischen Passfähigkeit zu einer Stelle und bestimmten Charakteristika aus Bewerbungs- oder Personalunterlagen herzustellen, und wären in diesen Beschäftigungsverhältnissen Frauen unterrepräsentiert, etwa weil ihnen der Zugang zu diesen Beschäftigungsverhältnissen verwehrt wurde, dann ist dieses Missverhältnis auch in der Zusammensetzung der Merkmale des Modells zu erwarten (Calders & Žliobaitė 2013: 50). Dieses Problem ist im Beispiel 1 (S. 34) des Systems zur Personalsuche des Unternehmens Amazon zu finden, bei dem die Modelle mit zu Lasten des Frauenanteils verzerrten Datensätzen trainiert wurden. Oder werden zur Modellbildung historische Datensätze genutzt, können die Daten zwar frühere Ungleichheiten korrekt wiedergeben. Wenn sich aber die Verhältnisse in der Zwischenzeit verändert haben, dann bilden die Modelle die aktuellen Verhältnisse nicht korrekt ab. Beispielsweise würde ein Modell, dass auf historischen Daten zu Einkommensverhältnissen zwischen Frauen und Männern beruht, und z. B. für die zielgerichtete Werbung an wirtschaftlich vielversprechende Kund*innen genutzt werden soll, die historischen Ungleichheiten in den Einkommensverhältnissen korrekt abbilden. Bei sich mittlerweile veränderten Berufstätigkeiten und Einkommensrelationen würden die erzeugte Auswahl und Relationen der Merkmale aber in dem Modell die aktuellen Verhältnisse nicht korrekt wiedergeben (Calders & Žliobaitė 2013: 50f.).

Viele Beispiele in Kapitel 4 mit Analysen oder Verwendungen von Verfahren des maschinellen Lernens oder Data-Minings zeigen Diskriminierungsrisiken aufgrund von verzerrten (Trainings-)Datensätzen, die (vorausgegangene) Ungleichbehandlungen, Stereotypen und Diskriminierungen abbilden.[77] Das Beispiel der stereotypenverstärkenden

[77] Siehe Beispiele 1, 8, 34, 35, 37, 38, 39, 40, 41, 45 und 46.

Ergebnisse der Bildsuchmaschine (Beispiel 15) zeigt auch, dass es zu qualitativen Verzerrungen in den Datensätzen kommen kann, wenn die Bilddaten stereotypbelastet gekennzeichnet sind. Die spezielle Untersuchung zu den Datengrundlagen von Systemen der vorausschauenden Polizeiarbeit in den USA (Beispiel 34) verdeutlicht, dass die Abbildung von irregulären Polizeipraktiken in den Datensätzen die Legitimität der gesamten Systeme in Frage stellen könnte.

(2) Des Weiteren können Daten aus der Nutzung von Diensten oder Produkten stammen, die von bestimmten Personengruppen weniger oder gar nicht genutzt werden (Lerman 2013). Dies kann beispielsweise der Fall bei Auswertungen der Nutzung von bestimmten IT-Diensten sein (z. B. Onlinedienste oder Breitbandnutzung), insbesondere bei Datensätzen, die aus sozialen Onlinenetzwerken stammen (Hargittai 2015). (3) Ebenso kann aus Kostenüberlegungen die Datenerhebung bewusst beschränkt oder auf kostengünstig verfügbare, aber ungeeignete Datensätze zurückgegriffen werden. (4) Auch aus Gründen des Datenschutzes oder des Schutzes der Privatsphäre können bestimmte Erhebungen nicht erfolgen (Calders & Žliobaitė 2013: 52f.). (5) Zu einer Überrepräsentation einer Bevölkerungsgruppe kann es kommen, wenn sich Aktivitäten, mit denen eine Datenerfassung verbunden ist, überproportional auf eine bestimmte Personengruppe konzentriert und auf diese Weise überproportional viele Erhebungsgegenstände erfasst werden (Calders & Žliobaitė 2013: 51). Letzteres liegt z. B. bei Polizeiaktivitäten vor, die als Ergebnis von Datenauswertungen nochmals gesteigert in bereits stark kontrollierten Gegenden erfolgen und so zu Verstärkungseffekten führen können (Beispiel 32, S. 62).

Das einfache Weglassen von geschützten Merkmalen mit der Absicht, Diskriminierung zu vermeiden, kann selbst Diskriminierungsrisiken hervorrufen (Calders & Žliobaitė 2013; Žliobaitė & Custers 2016). Modelle, die scheinbar „neutrale“ Merkmale bzw. **Proxies** anstelle von geschützten Merkmalen verwenden, können ebenso zu Risiken der mittelbaren Diskriminierung führen, wenn zwischen den Proxies und den geschützten Merkmalen eine Korrelation besteht (Barocas & Selbst 2016: 720–722). Beispielsweise kann bei einem Modell, das die Kreditwürdigkeit bestimmen soll, zwischen Wohnort und der ethnischen Zugehörigkeit ein Zusammenhang bestehen.

Selbst wenn man das geschützte Merkmal „Ethnie“ weglassen würde, kann gegebenenfalls aus dem Wohnort auf die ethnische Herkunft geschlossen

werden, wenn der Wohnort mehrheitlich von einer Personengruppe gleicher ethnischer Herkunft bewohnt ist. In diesen Situationen würde auch die Verwendung „neutraler" Faktoren eigentlich geschützte Gruppierungen benachteiligen. Generell ergibt sich das Problem, wenn die Merkmale nicht unabhängig voneinander sind, da dann nicht bestimmt werden kann, welches Merkmal mit welchem Ausmaß zum Modell beiträgt (Calders & Žliobaitė 2013: 47). Verdeutlicht wird das Problem auch am Beispiel 8 (S. 40), bei dem das Weglassen von Geschlechtsindikatoren keine Verbesserung bei Ungleichheiten in Berufsgruppenklassifizierungen brachte, und nach Meinung der Autor*innen die Verwendung von maschinellen Lernverfahren die Geschlechterunterschiede weiter verstärken kann.

Insgesamt können mit Verfahren des Data-Minings und des maschinellen Lernens mehr Variablenarten bzw. Dimensionen verarbeitet werden im Vergleich zu „klassischen" statistischen Verfahren. Dadurch steigt auch das Risiko von (unbemerkten) Korrelationen zu geschützten Merkmalen. Des Weiteren können sensible personenbezogene Daten gerade dann erforderlich sein, um gegebenenfalls in einem aggregierten Zustand mithilfe von statistischen Untersuchungen Diskriminierungen identifizieren und ausgleichen zu können (z. B. FRA 2018: 9f. zum Merkmal ethnische Herkunft).

Ähnliche Probleme können auftreten, wenn Modelle bzw. in ihnen enthaltende Systeme in anderen Lebensbereichen oder Kontexten angewandt werden, und die ursprüngliche Grundgesamtheit, die zur Modellbildung verwendet wurde, nicht zu der Grundgesamtheit des neuen Anwendungsbereichs passt[78]. Das stellt grundsätzliche Fragen an die Übertragbarkeit und (kommerzielle) Handelbarkeit derartiger Systeme. So weisen Schweighofer u. a. darauf hin, dass das COMPAS-System (Beispiel 35) ursprünglich zur Unterstützung bei Bewährungsentscheidungen gedacht war, aber in einigen Bundesstaaten auch zu Strafurteilen eingesetzt wird. (Schweighofer u. a. 2018: 39)

5.1.3 Risiken bei Onlineplattformen

Onlineplattformen unterscheiden sich von „einfachen" Webseiten dadurch, dass sie verschiedene Akteur*innen zusammenbringen und ihnen eine

[78] Das kann beispielsweise zu einer schlechteren Erkennung von bestimmten Personengruppen bei KI-basierten Systemen der Gesichts- oder Spracherkennung führen. Siehe dazu die Beispiele 37 (S. 68), 42 (S. 72) oder 46 (S. 74).

Möglichkeit des sozialen und wirtschaftlichen Austausches bieten, indem sie die Kommunikationsmöglichkeiten (z. B. bei sozialen Onlinenetzwerken) oder die Handelsfunktions- oder das „match making" (z. B. bei Onlineplattformen für die Arbeitsvermittlung) zur Verfügung stellen. In vielen Beispielen[79] konnten Ungleichbehandlungen und Diskriminierungen bei Onlineplattformen nachgewiesen werden. Einer der Ursachen ist die Ermöglichung der Bewertungen und Selektionen von Nutzenden durch andere Nutzende. Treffend dazu: „Full of salient pictures and social profiles, these platforms make it easy to discriminate [...]" (Edelman & Luca 2014: 10).[80] Algorithmen können dabei die Analyse personenbezogener Daten und die Kategorisierung oder das Ranking der Individuen durchführen oder administrieren die Informationen, die für die Teilnehmenden der Interaktionen und Transaktionen einstellbar oder sichtbar und nutzbar sind.

Darüber hinaus haben Algorithmen noch andere diskriminierungsrelevante Funktionen im Zusammenhang mit Onlineplattformen. Im Beispiel 2 (S. 35) wird vermutet, dass auf einer Onlineplattform der Arbeitsvermittlung auch die Rangfolge von Suchergebnissen auf Basis der Bewertungen und Einstufungen der Arbeitesanbietenden durch die Nachfragenden gebildet wird. Da bereits die Bewertungen nach Geschlecht und ethnischer Herkunft verzerrt waren, kommt es zu einer Fortsetzung der Verzerrungen in den Rangfolgen der Suchergebnisse. Auf Onlineplattformen aufgefunden zu werden, kann erhebliche wirtschaftliche Folgen für den Zugang zu Arbeitsstellen und Einkommensmöglichkeiten haben. Im Beispiel 27 (S. 55) wird verdeutlicht, dass sich verzerrte Restaurantbewertungen in Ungleichbehandlungen in der algorithmenbasierten Risikovorhersage der staatlichen Kontrolle fortsetzen können. Die algorithmische Fortsetzung von Voreingenommenheiten und Diskriminierungen kann so zu gesellschaftlichen Akkumulations- und Verstärkungsrisiken führen (siehe Abschnitt 5.2.2, S. 89).

---

[79] Siehe Beispiele 2, 3, 4, 5, 6, 7, 9, 10, 17, 18, 19, 20 und 26.

[80] Beispielsweise weisen Forschende in einer wissenschaftlichen Studie Ungleichbehandlungen beim sogenannten „peer-to-peer lending", d. h. der Vergabe von Krediten über Onlineplattformen, am Beispiel von Daten zu Transaktionen des Onlinemarktes Prosper.com nach. Sie zeigen unter anderem, dass Kreditsuchende, die in den Profilen Personenbilder mit Afroamerikanischem Aussehen haben, mit einer geringeren Wahrscheinlichkeit von 25 bis 30 Prozent eine Finanzierung erhalten haben. Da jedoch auch Begünstigungen für Afroamerikanische Kreditsuchende bei den Nettoverzinsungen nachgewiesen werden konnten, konnten die Autoren keine eindeutigen Schlussfolgerungen über das Vorliegen einer präferenzbedingten oder statistischen Diskriminierung ziehen. Allerdings scheinen in diesem Beispiel nicht Algorithmen die Ursache von Diskriminierungen zu sein (Pope & Sydnor 2011).

Des Weiteren sind bestimmte Onlineplattformen, insbesondere Suchmaschinen oder „soziale“ Onlinenetzwerke, Unternehmen der Aufmerksamkeitserzeugung und die Haupteinnahmequelle sind Erlöse aus der personen- oder gruppengerichteten Werbung. Dabei ermöglichen Algorithmen auch die Bepreisungs- und Marktmechanismen (z. B. Auktionsmechanismen) für die Werbungschaltung und Kundenselektion. Die Beispiele 17 (S. 46), 18 (S. 47), 19 (S. 47) und 20 (S. 48) zeigen, dass diese algorithmischen Vermarktungsmechanismen Ursachen von Ungleichbehandlungen und Diskriminierungen sein können. Algorithmen können selbst dann zu Diskriminierungen führen, wenn Nutzende eigentlich keine diskriminierenden Einstellungen vorgenommen haben, wie beispielsweise im Fall 20 (S. 48), bei dem die Auswahl der Zielgruppen von Werbung algorithmenbasiert stattfindet.

5.1.4 Absichtliche Diskriminierung und Verschleierung in und durch Computersysteme

Alle genannten Risikoquellen könnten auch durch Entwickelnde und Anwendende genutzt werden, um absichtliche Diskriminierung zu verschleiern, d. h., die Datensätze könnten wissentlich mit Verzerrungen ausgewählt sein oder es könnten gerade diejenigen verwendet werden, die bekanntlich Ungleichbehandlungen aus der Vergangenheit abbilden. Ebenso könnten bewusst Modelle mit Merkmalen gewählt werden, die einzelne diskriminierungsanfällige Gruppierungen nicht richtig erkennen (Barocas & Selbst 2016: 692; Dwork & Mulligan 2013; Kim 2016). Da Data-Mining und maschinelles Lernen es erlauben, geschützte Merkmale aus Datensätzen, die sie gar nicht enthalten, abzuleiten, können Entscheidende potenziell auch mit Datensätzen ohne geschützte Merkmale diskriminieren. Falls sie zur Rechenschaft gezogen würden, könnten sie darauf verweisen, ausschließlich „nicht geschützte“ Merkmale verwendet zu haben. In Bereichen, in denen der Nachweis von mittelbaren Diskriminierungen ohnehin schwierig ist, wie z. B. im Arbeitsbereich, könnten derartige Praktiken den Nachweis noch weiter erschweren (Barocas & Selbst 2016: 692f.).

5.1.5 Unzureichende Anreize zur Revision oder Abschaffung

Kim verdeutlicht anhand des Arbeitsbereichs, dass wettbewerbliches Vorgehen, marktwirtschaftliche Interessen und Effizienzorientierung der Anwendenden keine ausreichenden Anreize liefern können, um die Anwen-

dung von Analyse- und Entscheidungssystemen zu hinterfragen und gegebenenfalls zu verändern oder abzuschaffen, selbst wenn diese zu Diskriminierungen führen. Als Gründe dafür nennt sie: (a) Systeme mit Diskriminierungsrisiko können immer noch „genau" genug sein, sodass deren Anwendung nicht hinterfragt werden, (b) Rückkopplungseffekte können die „Genauigkeit" des Systems stabilisieren und (c) Systeme können sogar deshalb effizient sein, weil sie diskriminieren (Kim 2016: 892-897).

So kann beispielsweise die Genauigkeit bei einem Kriterium dazu führen, dass die Überprüfung der Tauglichkeit des gesamten Systems vernachlässigt wird. Auch Analyseverfahren des maschinellen Lernens haben nur eine begrenzte Datenmenge mit begrenzten Merkmalen. Daher bilden auch die resultierenden Modelle nur einen begrenzten Betrachtungsbereich der für Entscheidungen potenziell relevanten Kriterien ab. Insbesondere je „erfolgreicher", im Sinne der Genauigkeit bei einem Kriterium, ein Analyse- und Entscheidungssystem ist, desto mehr kann die Fokussierung auf dieses Kriterium erfolgen. Die Anwendenden haben dann keinen Anreiz mehr, Schlussfolgerungen und Mechanismen des Systems zu hinterfragen, auch wenn dieses systematisch Diskriminierungen verursacht. Dann werden sie vor allem auch nicht hinterfragen, ob nicht ganz andere entscheidungsrelevante Kriterien relevant sind und beachtet werden müssten (Kim 2016: 894f.).

5.2 Gesellschaftliche Risiken von algorithmenbasierten Differenzierungen

Auch wenn Diskriminierungsrisiken der Algorithmen und Daten soweit wie möglich vermieden werden würden, können gesellschaftliche Diskriminierungsrisiken aus der Verwendung von algorithmenbasierten Differenzierungsverfahren und automatisierten Entscheidungssystemen an sich resultieren. Sie sind nicht durch technische Lösungen zu beseitigen, sondern bedürfen der gesellschaftlichen Lösung durch geeignete Formen der politischen Handhabung und gegebenenfalls Regulation. Derartige Risiken werden auch als unintendierte Folgen, negative Konsequenzen, soziale Kosten oder Externalitäten bezeichnet (Gandy Jr. 2010: 36–39), weil sie bei Entscheidungen über den Einsatz von Differenzierungsverfahren durch die Anwendenden nicht ausreichend berücksichtigt werden und bei Betroffenen oder Dritten anfallen. Allerdings können die gesellschaftlichen Risiken drastisch erhöht und verstärkt werden, wenn zusätzliche systematische Fehler bei Algorithmen und Datensätzen vorliegen.

5.2.1 Gruppenzugehörigkeit und Generalisierungsunrecht

Wie dargestellt, liegt bei algorithmen- und datenbasierten Differenzierungsentscheidungen oft das Phänomen der statistischen Diskriminierung vor, da Differenzierungen entlang von Ersatzinformationen stattfinden, die entweder die geschützten Merkmale sind oder Korrelationen zu ihnen aufweisen (siehe Abschnitt 3.3). Charakteristisch dabei ist, dass die Ersatzinformationen häufig durch Analysen von Daten zu Personengruppen erzeugt werden. Dies ist weitgehend bei den Verfahren des Data-Minings und des maschinellen Lernens der Fall. Dabei stellt sich die Frage, ob es gerecht ist, dass Individuen mithilfe von Daten über andere Personen bewertet werden, d. h. auf Basis von Gruppen, zu denen die betroffenen Individuen selbst nicht unbedingt gehören müssen (z. B. Eckhouse u. a. 2019: 198f.).

Das grundsätzliche Problem von statistischen Auswertungen mit dem Gebrauch von aggregierenden Parametern bzw. Messgrößen ist, dass sie lediglich Aussagen über Merkmale einer bestimmten Population bzw. einer Aggregierung von Personen sind. Sie sollten eigentlich nur auf einen Aspekt der Population Bezug nehmen, jedoch werden die Parameter oft so gebraucht, als ob jedes der Mitglieder dieser Gruppierung das Merkmal aufweisen würde. Dadurch erhalten die Parameter Eigenschaften von Stereotypen (Gandy Jr. 2010: 34).

Bei Algorithmen zur Vorhersage, wie insbesondere bei der Berechnung von Risikoscores, ist das Vorhersageergebnis nicht das wahrscheinliche künftige Verhalten oder der künftige Zustand der betroffenen Person, sondern meistens eine Fortschreibung der vorhergehenden Bewertungen von anderen Personen durch andere Personen. Besonders deutlich wird dies am Beispiel 22 zu Risikoscores beim finnischen Fall der Onlinekreditvergabe und dem Beispiel 35 zum COMPAS System mit Risikoscores über Rückfallwahrscheinlichkeiten.[81] In den Gerichtsurteilen wurde darauf hingewiesen, dass die Bewertungen von Individuen allein auf Basis von statistischen Auswertungen von Daten, die wiederum Bewertungen von anderen Personen abbilden bzw. Gruppendaten sind, rechtlich problematisch sind. Dies ist insbesondere der Fall, wenn andere Informationen zu den Individuen bei ihrer Bewertung nicht herangezogen werden. Auch die Fallbeispiele zu

[81] Dies gilt auch für Risikoscores im Sozialbereich, was Beispiel 28 verdeutlicht.

Bewertungsdaten beim Fahrdienst Uber (Beispiel 26, S. 54), Risikobestimmung mit Restaurantbewertungen (Beispiel 27, S. 55) oder Risikobestimmung mit quasi Nachbarschaftsbewertungen im Kindesschutz (Beispiel 28, S. 55) verdeutlichen das Problem.

Es kommt zu Risiken von Fehlinterpretationen und Fehlschlüssen[82] auf Basis der (vermeintlichen) Gruppenzugehörigkeit (oder der Gruppenzugehörigkeit nach Regionen) (Schauer 2018; Lippert-Rasmussen 2007; Kamp & Weichert 2005: 51; SVRV 2018: 48; Zweig, Fischer & Lischka 2018: 25). Sie sind besonders problematisch, wenn es zur Stigmatisierung durch Falschzuschreibungen von negativen persönlichen Eigenschaften kommt, wie z. B. eine statistisch ermittelte „Unzuverlässigkeit" bei Kreditentscheidungen (Britz 2008: 124) oder die „Verbrechensneigung" bei Urteilen zur Inhaftierung (Eckhouse u. a. 2019).

Solche statistischen und algorithmenbasierten Entscheidungen werden dann nicht dem Einzelfall gerecht. Beeinträchtigungen resultieren daraus, dass über eine oder mehrere Personen, die ein bestimmtes Merkmal aufweisen, eine Annahme getroffen wird, die zwar auf die Mehrzahl der Merkmalstragenden zutreffen kann, nicht aber auf die konkrete Person oder die konkreten Personen im Einzelfall zutreffen muss (nach Britz 2008: 120f.). Aus verfassungsrechtlicher Sicht kann im Hinblick auf die Gleichheitssätze des Grundgesetzes (Art. 3 GG)[83] bei der statistischen Diskriminierung ein **Generalisierungsunrecht** auftreten, beispielsweise wenn „atypische" Personen, die beispielsweise nach einem Ersatzmerkmal (z. B. Alter) von bestimmten beruflichen Tätigkeiten ausgeschlossen werden, diese aber eigentlich noch ausüben könnten (Britz 2008: 2-11). Ebenso kann von einer Unvereinbarkeit mit der Einzelfallgerechtigkeit gesprochen werden, denn:

„In Fällen statistischer Diskriminierung wird ein Mensch wegen eines bestimmten (Stellvertreter-)Merkmals anhand stereotyper Personenvorstellungen beurteilt und behandelt, ohne dass seine tatsächlichen Eigenschaften gewürdigt würden." (Britz 2008: 12)

---

[82] An dieser Stelle stehen die in Abschnitt 5.1 beschriebene Risiken in einem direkten Zusammenhang.

[83] „(1) Alle Menschen sind vor dem Gesetz gleich. (2) Männer und Frauen sind gleichberechtigt. Der Staat fördert die tatsächliche Durchsetzung der Gleichberechtigung von Frauen und Männern und wirkt auf die Beseitigung bestehender Nachteile hin. (3) Niemand darf wegen seines Geschlechtes, seiner Abstammung, seiner Rasse, seiner Sprache, seiner Heimat und Herkunft, seines Glaubens, seiner religiösen oder politischen Anschauungen benachteiligt oder bevorzugt werden. Niemand darf wegen seiner Behinderung benachteiligt werden." Art. 3 GG.

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 Jr. 2010: 36f.). Diese können kaum durch technische oder organisatorische Verbesserungen gelöst werden, sondern bedürfen der **gesellschaftlichen Abwägung und Festlegungen** in politischen und rechtssetzenden Prozessen.[84] Der allgemein verbindliche Ausgleich soll eigentlich durch das Recht, insbesondere durch das AGG geschehen. Aufgrund der relativ unkonkreten, generalklauselartigen Ausnahmeregelungen des AGG (Britz 2008: 72) und immer dann, wenn neue Formen und Anwendungen des Typs der statistischen Diskriminierung auftreten, muss die Frage der Legitimität immer wieder neu für die jeweilige Einzelsituation gestellt werden. Hierzu wären algorithmische Verfahren der üblichen Prüfung nach der Verhältnismäßigkeit auf ihren legitimen Zweck, ihre Eignung, Erforderlichkeit und Angemessenheit zu unterziehen (Britz 2008: 151–179).

Dabei sind die jeweils besonderen Eigenschaften der jeweiligen Differenzierungssituationen zu beachten, insbesondere ob Differenzierungen auf Gruppendaten oder auf Erfassungen des individuellen Verhaltens beruhen, aber auch der Grad der Genauigkeit der algorithmischen Ergebnisse bzw. die Fehlerraten und sonstigen technischen Risiken sowie das Wissen über ursächliche Zusammenhänge[85] oder das Ausmaß bzw. die Schwere der Benachteiligung durch die Fehlschlüsse bei Differenzierungsentscheidungen (z. B. Ablehnung eines Kredits, Nichteinstellung in ein Arbeitsverhältnis, höhere Versicherungstarife).

So bemisst sich der Grad der Benachteiligung etwa daran, welches Gut den Benachteiligten vorenthalten wird und wie die Schwere der Einschränkung der Entfaltungsmöglichkeiten der betroffenen Person beurteilt wird (Britz 2008: 125). Beispielsweise sind Entscheidungen über Gefängnisstrafen und Freiheitsentzug sowie Entscheidungen über die Chancenverteilung zur Persönlichkeitsentwicklung (z. B. über Bildungs- oder Berufszugänge) anders zu beurteilen als Entscheidungen über die Auswahl für zielgerichtete Werbung für Konsumgüter. Dennoch werden auch aus Kos-

[84] Siehe auch Abschnitt 6.4.2, ab S. 139.
[85] Siehe Schauer (2018: 46).

tengründen algorithmenbasierte Systeme der Entscheidungsunterstützung bei Entscheidungen mit gravierenden Konsequenzen für die Persönlichkeitsentfaltung in Gerichtsverfahren eingesetzt (Eckhouse u. a. 2019).

5.2.2 Akkumulations- und Verstärkungseffekte

Risiken und negative Effekte einer wirtschaftlich rationalen Differenzierung können sich akkumulieren und verstärken. Insgesamt kann es zu kumulativen Benachteiligungen im Sinne der Einschränkung von Lebensentwicklungs- und Entfaltungschancen, der Einkommenssicherung, dem Grad der politischen Involvierung sowie der Durchsetzung von Gerechtigkeit im Rechtssystem kommen (Gandy Jr. 2010: 37). Das sind keine grundsätzlich neuen, erst durch Algorithmen verursachte Risiken, aber die Diskriminierungsrisiken durch die Verwendung von Algorithmen können genauso zu Akkumulations- und Verstärkungseffekten beitragen. Risiken und Effekte treten insbesondere dann auf, wenn ein diskriminierungsanfälliges Merkmal als Ersatzinformation für die Differenzierungsentscheidungen herangezogen wird, und es zu einer sich gegenseitig verstärkenden Wirkung von Stigmatisierung, Beeinträchtigung der Selbstdarstellung[86], Benachteiligung durch Falschzuordnung und den daraus erschwerten Zugängen zu Gütern, die der Persönlichkeitsentfaltung dienen, kommt.

Ein weiterer Verstärkungseffekt ergibt sich, wenn bestehende Ungleichbehandlungen in der Bevölkerung wahrgenommen und dadurch den Betroffenen Anreize genommen werden, z. B. sich weiterzuqualifizieren. Werden künftige Diskriminierungen antizipiert, können die Investitionen in Qualifizierungen nicht mehr lohnend erscheinen (Britz 2008: 126f.). Dies wird durch das Beispiel 15 (S. 45) veranschaulicht, bei dem das Risiko diskutiert wird, dass durch die stereotypenverstärkende Bildsuchergebnisse zu Berufsgruppen das Karrierestreben der betroffenen Personengruppen beeinträchtigt werden kann. Im Allgemeinen dürften derartige Verstärkungseffekte vor allem von Systemen ausgehen, die Ungleichgewichte in der Repräsentation von Personengruppen bzw. Repräsentationsrisiken („reprentational harms“) aufweisen (z. B. Tolan 2018: 17).

[86] Siehe dazu den Abschnitt 5.2.5, ab S. 93.

Andere Typen von algorithmenbasierten Akkumulations- und Verstärkungseffekten resultieren dann, wenn Algorithmen auf Basis von bereits verzerrten oder diskriminierenden Bewertungen von Menschen durch Menschen gebildet (und kontinuierlich angepasst) und für andere Funktionen verwendet werden, wie sich dies bei den Beispielen 28 (S. 55) und 32 (S. 62) zur der Bildung von Risikoscores auf Basis von menschlichen Bewertungen bzw. Beurteilungspraktiken zeigt. Ebenso wird dies am Beispiel 27 (S. 55) deutlicht, bei dem verzerrte Kundenbewertungen von verschiedenen ethnienorientierten Restaurants als Basis für Algorithmen von Vorhersagesystemen in der staatlichen Aufsicht verwendet werden.

5.2.3 Differenzierungen gegen gesellschaftspolitische Vorstellungen

Auch wenn Effizienzgewinne durch den Gebrauch von Differenzierungen mit Ersatzinformationen vorliegen würden, kann es gesellschaftspolitisch gewünscht sein, dass auf Differenzierungen verzichtet wird, um Ziele der Gleichheit und gleichen Behandlung zu verwirklichen, wie etwa die Gleichheit vor Gericht, beim gleichen Zugang zu Infrastrukturen, gleiche Bildungs- und Aufstiegschancen und gleiche Behandlung mit Respekt und Würde. Ebenso kann eine mögliche Differenzierung unterbunden werden, wenn vormals diskriminierte oder besonders diskriminierungsanfällige Gruppierungen geschützt und gefördert werden sollen (Schauer 2018: 50).

Konkret kann eine ökonomisch rationale Differenzierung bzw. Benachteiligung bestimmter Merkmalsträger abgelehnt werden, wenn (a) der Wunsch nach Kompensation für vergangenes Diskriminierungsunrecht bzw. struktureller Benachteiligung bestimmter Merkmalsträger besteht, gerade um Akkumulations- und Verstärkungseffekte zu durchbrechen. Des Weiteren sollte auf eine Differenzierung verzichtet werden, wenn (b) eine Differenzierung für die Mitglieder einer strukturell benachteiligten Gruppe den Zugang zu Gütern, Ressourcen und Positionen erschweren würde, die sie gerade zur Überwindung des benachteiligten Gruppenstatus benötigen würde (z. B. Zugang zu Beschäftigungsverhältnissen oder Krediten). Ferner kann eine ökonomisch rationale Differenzierung abgelehnt werden, um (c) eine expandierende Stereotypisierung zu vermeiden. Denn ist eine Differenzierung mit der Zuschreibung einer negativen Eigenschaft verbunden, kann dies zu expandierender Stereotypisierung führen, wenn vor allem eine Gruppierung betroffen ist, die ohnehin mit negativen Stereotypen konfrontiert ist. Schließlich können (d) weitere Gründe, wie gesundheits- oder

sozialpolitische Ziele, gegen eine ökonomisch rationale Differenzierung, wie z. B. bei Kranken- oder Kfz-Versicherungstarifen, sprechen (Britz 2008: 127–130).

Gesellschaftliche Risiken könnten in Zukunft entstehen, wenn gesellschaftliche Abwägungen über durch Algorithmen ermöglichte oder verbesserte und wirtschaftlich sinnvoll erscheinende Differenzierung einseitig zu Gunsten von Effizienzbestrebungen und zu Lasten von Gleichheitsbestrebungen oder sozialpolitischen Zielen laufen. Die Gefahr kann sich erhöhen, wenn Kostensenkungen durch Automatisierungen bei algorithmen- und datenbasierten Differenzierungen die Differenzierungen in vielen neuen Anwendungsbereichen erst ermöglichen und andere Regelungsformen verdrängen.

5.2.4 Behandlung als ein bloßes Mittel und psychologische Distanzierung

Die Verwendung algorithmenbasierter Differenzierung verstärkt das Risiko, dass Menschen nicht mehr als Individuen bzw. in Anerkennung ihrer grundrechtlich verbrieften Menschenwürde und ihrer einmaligen individuellen Subjektqualität[87] wahrgenommen, sondern nur noch **als bloßes Objekt bzw. Mittel** gesehen und behandelt werden. Zwar ist es nach dem Instrumentalisierungsverbot bzw. der Objektformel möglich, andere Menschen auch als Mittel für seine Zwecke zu behandeln, aber es ist nach diesem moralischem Prinzip untersagt, sie ausschließlich als bloßes Mittel zu benutzen beziehungsweise sie als bloßes Mittel zu eigenen Zwecken herabzuwürdigen. Menschen als bloßes Mittel zu behandeln, bedeutet, wenn man sie in einer Weise behandelt, der sie nicht **zustimmen** können. Das kann z. B. bei einem falschen Versprechen der Fall sein, wenn die Betroffenen nicht wissen, was man mit ihnen tatsächlich vorhat. Ebenso können sie einer Behandlung nicht zustimmen, weil sie keinen Grund dazu haben, oder wenn es irrational wäre, zuzustimmen. Andere Personen in ihrer Würde zu achten, heißt dann, sie in einer Weise zu behandeln, die ihnen die

[87] Vgl. Wiegerling (2016) und Hänold (2018: 130f.). Siehe auch Härtel (2019: 60), die mit dem Verweis auf den kategorischen Imperativ von Immanuel Kant „Handle so, dass Du die Menschheit in Deiner Person und in der Person jedes anderen jederzeit zugleich als Zweck, niemals bloß als Mittel brauchst." Kant (1786/1977: 60) die Forderung stellt, dass das Grundprinzip des Schutzes der Menschenwürde alle Regelungen zur digitalen Transformation durchwirken solle.

Möglichkeit gibt, zu dem, was man mit ihnen tut, vernünftigerweise zuzustimmen oder die Behandlung ablehnen zu können (Schaber 2012: 40–42).

Bei wirtschaftlichen Anwendungen wollen Anwendende das Verhalten und die Zustände der Betroffenen erfassen und analysieren, um in erster Linie den monetären Wert aus den Beziehungen zur Kundschaft zu erlangen oder zu steigern und nicht, um die eigentlichen Gründe für das Verhalten herauszufinden (Yeung 2018: 30). Algorithmische Datenanalysen des Data-Minings, der Big-Data-Analytik oder des maschinellen Lernens erzeugen typischerweise Korrelationen und keine Kausalzusammenhänge. Dadurch fehlt den Entscheidenden die Grundlage, den Betroffenen die Gründe ihrer Entscheidung, z. B. wenn sie aussortiert werden, hinreichend zu erläutern. Dadurch haben die Betroffenen nicht die Möglichkeit, der Behandlung, der sie unterzogen werden, zuzustimmen oder sie abzulehnen. Dieses Risiko wurde insbesondere im Beispiel 22 (S. 50) deutlich, bei dem durch die Verwendung des Systems dem Betroffenen nicht ausreichend die Ablehnung des Kredits erklärt werden konnte, was mit in die Begründung des Diskriminierungstatbestandes eingeflossen ist.

Zusätzlich besteht ein Risiko, dass algorithmenbasierte Entscheidungsverfahren eine **psychologische Distanzierung** der verantwortlichen Personen von den Entscheidungen und den Betroffenen Vorschub leisten. Das Risiko, das durch die Zwischenschaltung von Computern als „moralische Puffer“ sowie eine scheinbare Verlagerung der moralischen Verantwortung für die Entscheidungen auf Computer entsteht, wurde zwar bisher hauptsächlich für autonome Waffensysteme diskutiert (z. B. Cummings 2004b; Brundage u. a. 2018: 17), es kann aber für alle halb- und vollautomatisierten Entscheidungsverfahren mit negativen Konsequenzen für die Betroffenen gesehen werden.

Das Datenschutzrecht hat zur Verminderung solcher Risiken das Verbot automatisierter Entscheidungen im Einzelfall entwickelt. Damit soll verhindert werden, dass nachteilige Differenzierungsentscheidungen ausschließlich durch automatisierte Verarbeitungsprozesse getroffen werden, und es soll erreicht werden, dass „[…] niemand zum bloßen Objekt einer allein auf Algorithmen basierenden Bewertung persönlicher Daten werden darf.“ (Scholz 2019: DSGVO Art. 22 Rn. 3; ebenso Martini 2018: DS-GVO Art. 22 Rn. 1). Im Abschnitt 6.2.3 (ab S. 114) wird das Verbot automatisierter Entscheidungen eingehender diskutiert und dessen Lücken aufgezeigt, die es fraglich erscheinen lassen, ob das eigentliche Schutzziel noch erreicht wird.

5.2.5 Gefährdung der freien Entfaltung der Persönlichkeit und des Rechts auf Selbstdarstellung

Neben den Gleichheitsrechten des Grundgesetzes (s. o.) sind durch das Phänomen der statistischen bzw. der algorithmen- und datenbasierten Diskriminierung auch die verfassungsrechtlich gewährten Persönlichkeitsrechte, insbesondere die durch Art. 2 Abs. 1 Grundgesetz[88] 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: 179f.; Fröhlich & Spiecker genannt Döhmann 2018).

„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: 124f.). 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: 180).

Dadurch wird den Betroffenen ihr **Recht auf Selbstdarstellung** genommen, das sich aus dem Recht auf freie Entfaltung der Persönlichkeit herleitet.[89] Nach Britz ist die Selbstdarstellung das Mittel des Individuums darauf Einfluss zu nehmen, welches Bild sich andere Menschen von ihm machen (Britz 2008: 179–207). Das Recht auf Selbstdarstellung dient der freien Entfaltung der Persönlichkeit auf zweifache Weise:

---

[88] „Jeder hat das Recht auf die freie Entfaltung seiner Persönlichkeit, soweit er nicht die Rechte anderer verletzt und nicht gegen die verfassungsmäßige Ordnung oder das Sittengesetz verstößt." Art. 2 Abs. 1 GG.

[89] Zur Ausgestaltung des Rechts auf freie Entfaltung der Persönlichkeit und dem Recht auf Selbstdarstellung siehe auch Britz (2007).

(1) Durch Selbstdarstellung kann erreicht werden, dass sich die anderen ein „günstiges" Bild von ihm machen und dadurch für die Wahrung seiner Entscheidungs- und Handlungsspielräume (**äußere Entfaltung**) günstige Entscheidungen treffen. Denn das Bild von einem Individuum ist immer dann entscheidend, wenn seine Handlungsspielräume von der Kooperationsbereitschaft anderer abhängen. Dann bestimmt das Bild von dem Individuum darüber, ob die Kooperationsbereitschaft überhaupt gegenüber dem Individuum gezeigt wird und ihm dadurch Handlungsspielräume eröffnet werden, z. B. ob ihm überhaupt ein Vertrag oder eine Mitgliedschaft angeboten wird. Oder das Individuum kann seine Handlungsfreiheit dadurch antizipativ selbst einschränken, wenn es nicht weiß, welche Fremdbilder von dem Individuum erzeugt werden. Hat man keinen Einfluss darauf, welche Daten und Informationen in das Fremdbild eingehen, kann bereits die bloße Antizipation von Fremdbildern prohibitiv wirken (Britz 2008: 190f.).

(2) Des Weiteren kann sich das Individuum im Sinne der **inneren Entfaltung** mit der „Selbstdarstellung einen hinreichenden Anteil am wechselseitigen Prozess der Konstituierung seiner Identität sichern, um sich selbst als freiwillig gewählte Persönlichkeit begreifen zu können." (Britz 2008: 195). Zwar findet Persönlichkeitsentwicklung in sozialen Kontexten immer in interaktiven Vorgängen von fremden Erwartungen und Zuschreibungen (Fremdbildern) einerseits und eigenen Selbstbildern, Vorstellungen und Wünschen andererseits statt, aber die Kerngewährleistung des Persönlichkeitsrechts ist es, „Mechanismen zur Verfügung zu stellen, die den Einzelnen so in die Vorgänge der Konstituierung von Persönlichkeit einbinden, dass er seine Persönlichkeit als frei gewählt begreifen kann [...]" (Britz 2008: 191).

Dabei geht es vor allem um den Schutz vor gesteigerten Formen der Fremdbestimmung, die die Persönlichkeitsentfaltung unfrei werden lassen. Dies geschieht dann, wenn einer Person intensive, d. h. durch eine besondere Qualität und Dichte gekennzeichnete Fremdbilder quasi „übergestülpt" werden und ihr dadurch den Raum für eigene Vorstellungen ihrer selbst nehmen. Dies kann (a) bei umfassenden datenbasierten Persönlichkeitsprofilen der Fall sein, bei denen die Seite der Bewertenden so umfassend über die Persönlichkeit des betroffenen Individuums informiert ist, dass für die eigene Rolleninterpretation in sozialen Kontexten keine Möglichkeit mehr bleibt. Diese Gefahr wird insbesondere auch durch das Recht auf informationelle Selbstbestimmung adressiert, bei dem die Selbstdarstellung als Bedingung der Entfaltung und Erhaltung von Persönlichkeit eben-

so gesehen wird (siehe auch Abschnitt 6.4.2).[90] Auch (b) durch den Gebrauch einer einzigen Ersatzvariable im Falle der statistischen Diskriminierung werden nicht die betroffenen Personen selbst wahrgenommen, sondern eine stereotyp konstruierte Persönlichkeit (Britz 2008: 193f.).

Zum Schutz der Persönlichkeitsentfaltung sind Diskriminierungsverbote 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: 200ff., 204).

5.2.6 Erzeugung von struktureller Überlegenheit

Algorithmische Analysemethoden, meist gestützt auf Verfahren der künstlichen Intelligenz, können zunehmend Persönlichkeitsmerkmale, Charaktereigenschaften und emotionale Zustände automatisiert identifizieren (siehe Abschnitt 2.2.2). Sie könnten dazu genutzt werden, die Angewiesenheit einer Person auf ein Produkt, einen Dienst, eine Ressource oder eine Position zu ermitteln und auszunutzen, wodurch die strukturelle Überlegenheit der Anbietenden erhöht würde. Der Effekt kann verstärkt werden, wenn die Anbietenden zusätzlich über die Zugangs- und Kontrollpunkte über umfangreiche Mengen personenbezogener Daten und Personenprofile verfügen, die Vorteile bei Verfahren des maschinellen Lernens bedeuten. Denn dort ist die Menge an Daten auch ausschlaggebend für die Qualität der erzeugten Modelle bzw. Algorithmen. Zusätzlich können auch Netzwerkeffekte[91], insbesondere bei Onlineplattformen oder IT-Systemen die strukturelle Überlegenheit der Anbietenden weiter erhöhen, denn Netzwerkeffekte bedeuten für die einzelnen Nutzenden einen (teils prohibitiv) hohen Wechselaufwand. Dadurch werden Wahl- und Ausweichmöglichkeiten verringert.

Dem Risiko der **strukturellen Unter- bzw. Überlegenheit** wird auch für private Verhältnisse zunehmend Beachtung geschenkt, so mit der Rechtsprechung des Bundesverfassungsgerichts zur strukturellen Unterlegenheit, u.a. mit der Entscheidung zu Bürgschaftsverträgen zur strukturell

[90] Vgl. Britz (2008:193), Hoffmann-Riem (1998), Trute (2003), (1998), Britz (2010) und Albers (2017).
[91] Zu Netzwerkeffekten siehe S. 4.

ungleichen Verhandlungsstärke (BVerfGE 89, 214 (1993)) oder der Stadionverbot-Entscheidung (BVerfGE 148, 267 (2018)). Sie können für Bereiche der Digitalisierung mit großen Machtasymmetrien relevant werden (Hoffmann-Riem 2017: 25; Schweighofer u. a. 2018: 78-80; Härtel 2019).

Zwar gilt weiterhin die verfassungsrechtlich geschützte Privatautonomie, nach der gehört es „[…] zur Freiheit jeder Person, nach eigenen Präferenzen darüber zu bestimmen, mit wem sie unter welchen Bedingungen Verträge abschließen will." (BVerfG 2018: Leitsätze). Aber nach der Entscheidung kann sich für spezifische Konstellationen der Geltungsbereich des Gleichheitssatzes des Art. 3 Abs. 1 GG auch auf privatwirtschaftliche Bereiche erstrecken. So dürfen Private ihre „[…] aus einem Monopol oder aus struktureller Überlegenheit – resultierende Entscheidungsmacht nicht dazu nutzen, bestimmte Personen ohne sachlichen Grund von einem solchen Ereignis auszuschließen." (BVerfG 2018: Rn. 41). Bei den adressierten Ereignissen handelt es sich um solche Veranstaltungen, die für die Betroffenen die „Teilhabe am gesellschaftlichen Leben" bedeuten (BVerfG 2018: Rn. 41). Daraus leiten Schweighofer u. a. (2018) die Kennzeichen von struktureller Überlegenheit ab, d. h. die Öffnung der Leistung für einen breiten Verkehr, die Angewiesenheit auf die Leistung und die einseitige Verfügungsmacht des anbietenden Unternehmens oder der anbietenden Person (ebd., S. 79).

Bei algorithmischen Entscheidungssystemen wird nicht per se eine strukturelle Überlegenheit vermutet, aber dann, „wenn die algorithmische Beurteilung von Personen genutzt wird, um die Angewiesenheit einer Person auf die Leistung […] zu erzeugen, zu verstärken oder auszunutzen." (Schweighofer u. a. 2018: 79) Die kann insbesondere dann vorliegen, wenn das algorithmische Verfahren dazu genutzt wird, um diejenigen Vertragspartner*innen zu identifizieren, die auf die Leistung angewiesen sind (ebd.). Nach Härtel könne dies bei dynamischen Preisen oder Preisdifferenzierungen, beim Kreditscoring oder bei Onlineplattformen mit Marktmacht vorliegen (Härtel 2019: 58). Daraus kann in den genannten Konstellationen ein Gleichbehandlungsgebot nach dem Gleichheitsgrundsatz des Art. 3 GG unter dem Gesichtspunkt der strukturellen Überlegenheit mit dem Hauptmerkmal der Angewiesenheit auf die Leistung resultieren. Es sind jedoch Voraussetzungen und Ausgestaltungen für die ausreichende Beachtung von struktureller Überlegenheit noch weitgehend unklar (Schweighofer u. a. 2018: 80).

6. Handlungsbedarfe und -optionen

Fast zeitgleich mit dem Erkennen von Diskriminierungsrisiken von Algorithmen hat die Suche nach Lösungsmöglichkeiten begonnen. Viele Empfehlungen oder Vorschläge zu Handlungsoptionen oder Instrumenten stammen aus dem internationalen Raum und sind nicht unmittelbar auf den hiesigen institutionellen Rahmen übertragbar. Beispielsweise wird zur Selbstprüfung durch die Entwickelnden und Anwendenden, u. a. auf Diskriminierungsrisiken, die Durchführung von Human Rights Impact Assessments (UN GA 2018; Yeung 2018: 65; Council of Europe 2019) oder Algorithmic Impact Assessments (Reisman u. a. 2018) vorgeschlagen. Bei diesen Instrumenten wäre aber das Verhältnis zum bestehenden Datenschutzrecht abzugleichen, insbesondere zu den Vorgaben zur Datenschutz-Folgenabschätzung (Art. 35 DSGVO), die bei der Verarbeitung von besonderen Kategorien personenbezogener Daten bzw. Daten mit besonders sensiblen bzw. diskriminierungsanfälligen Merkmalen (nach Art. 9 DSGVO) erforderlich ist, ebenso das Verhältnis zu den Vorgaben zu automatisierten Entscheidungen im Einzelfall (Art. 22 DSGVO), die ebenso dem Antidiskriminierungsrecht dienen sollen.

6.1 Transparenz und Nachweis von Diskriminierungen

Zur Transparenz von Algorithmen hat sich eine ausführliche Diskussion entwickelt[92]. Einige Autor*innen heben hervor, dass Algorithmen und Computersysteme durch Undurchsichtigkeit („opacity“) und Unverständlichkeit gekennzeichnet sind bzw. sogenannte „black box“-Eigenschaften aufweisen (z. B. Pasquale 2015; Castelvecchi 2016; Kitchin 2017). Quasi als Gegenmittel dazu wurde die Forderung nach Transparenz erhoben. Dabei kreist die Diskussion um Fragen über was, für welche Personen, für welchen Zweck und in welcher Form Transparenz geschaffen werden oder

[92] Teilweise mit unterschiedlichen Verständnissen des Begriffs „Transparenz“ von Offenlegung bis Erklärung. Siehe zur Diskussion beispielsweise Castelluccia & Le Métayer (2019: 26–30).

unterbleiben sollte (Citron & Pasquale 2014; Mittelstadt u.a. 2016; Ananny & Crawford 2018; de Laat 2017). Forderungen nach Schaffung von Transparenz können von der Erklärung der wichtigsten Funktionsweisen bis hin zur Offenlegung des Programmcodes bzw. der Möglichkeit der Einsichtnahme reichen. Zudem soll Transparenz sehr unterschiedliche Funktionen erfüllen, die von der Funktion als Informationsinstrument des Verbraucherschutzes bis hin zur Schaffung von Rechenschaft im Sinne der Algorithmic Accountability[93] gegenüber diversen Anspruchsgruppen reichen (Hacker & Petkova 2017).

In diesem Zusammenhang sind hinsichtlich der Ursachen von Intransparenz von Algorithmen und Computersystemen verschiedene Aspekte zu unterscheiden (nach Burrell 2016): (1) So ist die Intransparenz in vielen Fällen auf das Verhalten der Entwickelnden und Anwendenden zurückzuführen, die aus Gründen des Schutzes von Betriebs- und Geschäftsgeheimnissen, des Urheberrechtsschutzes, des Datenschutzes (wenn Computersysteme personenbezogene Daten von Dritten beinhalten) oder aus Vorsicht vor gezielten Verhaltensanpassungen durch die Betroffenen („gaming the system") die Offenlegung von Algorithmen, Programmstrukturen oder selbst der programmierten Entscheidungsregeln und -kriterien gegenüber Außenstehenden verweigern (de Laat 2017).

(2) Der Eindruck der Intransparenz kann auch durch unterschiedliche Fähigkeiten und Vorwissen des*r Betrachter*in entstehen. Ein Programmcode, der Algorithmen implementiert, ist nicht voraussetzungslos nachvollziehbar und in den wenigsten Fällen vollständig zu erfassen. Es bedarf Kenntnisse der Programmiersprache als eine grundlegende Voraussetzung, um Algorithmen „im Rohzustand" nachvollziehen zu können. Daher wäre eine Offenlegung des Programmcodes gegenüber Betroffenen ohne Vorkenntnisse wenig ergiebig, wohl aber kann eine Offenlegung für die Inspektion durch Fachkräfte sinnvoll sein (s.u. zum Testen von Softwaresystemen).

(3) Bei der Intransparenz durch die technischen Eigenschaften von Algorithmen und Softwaresystemen zeichnet sich ein differenziertes Bild ab. So wird darauf hingewiesen, dass die Intransparenz mit steigender Komplexität der Algorithmen und Softwaresysteme zunehmen kann (z.B. Yeung

[93] Zum Dachkonzept der „algorithmic accountability" mit US-amerikanischer Herkunft siehe Diakopoulos (2014), World Wide Web Foundation (2017), zur Einordnung in den Europäischen Kontext Busch (2018) und EDPS (2018).

2018: 15; Wischmeyer 2018: 47). Das schließt vor allem Algorithmen des maschinellen Lernens ein, ebenso adaptive bzw. dynamische Systeme, deren Regeln sich durch die kontinuierliche Analyse von Datenströmen ständig anpassen (Desai & Kroll 2017). Demgegenüber wird auch der Standpunkt vertreten, dass Algorithmen grundsätzlich verständliche technische Elemente sind und dass Undurchschaubarkeit vor allem durch die Interessen- und Machtstrukturen der Entwicklungsprozesse von Softwaresystemen entsteht (Kroll 2018). Dadurch, dass Entscheidungsregeln durch die Algorithmen programmiert sind und diese dafür spezifisch formuliert werden müssen, sehen Kleinberg u. a. (2019) die Regeln sogar als grundsätzlich nachvollziehbarer an als durch Menschen durchgesetzte Regeln. Denn mit Simulationen können die Ergebnisse der Entscheidungsregeln eindeutiger bestimmt werden als beim menschlichen Entscheiden.

6.1.1 Technische Optionen für Transparenz, Nachvollziehbarkeit und Diskriminierungsvermeidung

Derzeit erscheinen zahlreiche Vorschläge zur Schaffung von Transparenz und Nachvollziehbarkeit. Von den sich sehr schnell entwickelnden Forschungs- und Entwicklungsrichtungen, mit teilweise unklaren begrifflichen Abgrenzungen und weiten Überlappungen, können hier nur punktuelle Ausschnitte wiedergegeben werden.[94]

Für die **technische Analyse** von Algorithmen, vor allem des maschinellen Lernens und damit der Erzeugung von Erklärbarkeit, können verschiedene Ansätze unterschieden werden: (1) Ansätze mit „geöffneten" Systemen („white box approach"), bei denen es möglich ist, den Programmcode zu analysieren. (2) Davon sind Ansätze mit „geschlossenen" Systemen („black box approach") zu unterscheiden, bei denen das Verhalten eines Systems analysiert wird, ohne dass der Programmcode zur Kenntnis genommen wird. Dabei werden Erklärungen konstruiert, indem sowohl der Input als auch der Output beobachtet wird. (3) Des Weiteren kann noch der konstruktive Ansatz („constructive approach") unterschieden werden, der auf die Implementierung von Erklärbarkeit bereits bei der Entwicklung des Programmcodes abzielt (Castelluccia & Le Métayer 2019: 47–54). In diesem Zusammenhang

[94] Übersichten werden z. B. von Guidotti u. a. (2018), Schweighofer u. a. (2018), Castelluccia & Le Métayer (2019) oder Dosilovic, Brcic & Hlupic (2018) geliefert, mit jeweils anderen Einteilungen.

sind auch die Forschungsinitiativen der „sich selbsterklärenden" KI („explainable AI") zu sehen (z. B. Dosilovic, Brcic & Hlupic 2018). Beispielsweise stellen Ehsan u. a. (2019) ein KI-System vor, das seine Schritte in natürlicher Sprache erklären können soll („automated rationale generation").

Zur technischen Analyse heben Schweighofer u. a. die Möglichkeiten des Testens von Softwaresystemen hervor, die seit langem als Bestandteil des „System and Software Engineering" bestehen, ebenso als Standardverfahren vor allem der Qualitätssicherung dienen und dabei Ähnlichkeiten zum Auditing (s. u.) aufweisen. Beim Testen erhält ein Softwaresystem eine vorab definierte Eingabe und soll daraus eine Ausgabe erzeugen (Schweighofer u. a. 2018: 58-64).

Derzeit stehen einige Softwaresysteme zum Testen von Systemen des maschinellen Lernens und der automatisierten Entscheidungen größtenteils als Open-Source-Programme zur Verfügung (Sanchez-Monedero & Dencik 2018: 12f.). Beispiele sind das Themis System zum Testen von Fairness in Software (Galhotra, Brun & Meliou 2017), das FairTest Werkzeug zur Untersuchung von Zusammenhängen zwischen Ergebnissen von Anwendungen und sensiblen bzw. geschützten Merkmalen der Nutzenden (Tramèr u. a. 2017) oder das umfassende System „AI Fairness 360", das eine ganze Reihe von Werkzeugen integriert (Bellamy u. a. 2018).

Auf die besonderen Fähigkeiten, Entscheidungsregeln auf Diskriminierungen zu überprüfen, wenn sie in Algorithmen implementiert sind, weisen Kleinberg u. a. (2019). hin. Da relevante Entscheidungsregeln in den Algorithmen programmiert sind, seien Experimente und Simulationen möglich, mit denen die Auswirkungen der Entscheidungsregeln auf betroffene Personengruppen untersucht werden können, z. B. durch Variierung der Dateninputs oder der Entscheidungsregeln selbst. Voraussetzung für die Überprüfung ist jedoch der Zugang zu den Algorithmen und den Datensätzen. Im Gegensatz zu Algorithmen seien Menschen die „ultimative black box" (ebd., S. 10). Daher fordern die Autoren, dass man insbesondere die im Entwicklungs- und Anwendungsprozess durch Menschen getroffenen Entscheidungen, wie z. B. die Auswahl der Datensätze oder der Einflussvariablen, dokumentiert. Dadurch seien die Auswirkungen von Algorithmen nachvollziehbarer und Diskriminierungen könnten sogar bei Gerichtsverfahren, im Vergleich zum (konventionellen) Nachweis mit Statistiken, leichter nachgewiesen werden (Kleinberg u. a. 2019).

Mit Tests und Analysen verbunden sind technische Ansätze zur **Vermeidung von Diskriminierungen**, die bei der Gestaltung und Nutzung der Systeme ansetzen. Romei und Ruggieri unterscheiden dazu beim Data-Mining: (1) den naiven Ansatz des Weglassens von geschützten Merkmalen, bei denen auch sie jedoch auf dessen Probleme hinweisen.[95] Des Weiteren nennen sie (2) die kontrollierte Störung bzw. Veränderung des Trainingsdatensatzes („pre-processing"-Ansatz), (3) die Abänderung des Lernalgorithmus für die Klassifizierung während des Trainings („in-processing"-Ansatz), (4) die Veränderung des Modells zur Klassifizierung nach dem Trainieren („post-processing"-Ansatz) und (5) korrigierende Eingriffe bei der Anwendung des Vorhersagealgorithmus bzw. -modells (Romei & Ruggieri 2014: 622–624).[96] Im Beispiel 36 wird jedoch verdeutlicht, dass einige dieser Korrekturmöglichkeiten bei ML-Verfahren, die diskriminierende Risikoprognosen im Jugendstrafvollzug abgeben, nicht zu zufriedenstellenden Ergebnissen führen oder neue Diskriminierungen auslösen würden.

Zudem werden auch diskriminierungsvermindernde Algorithmen des maschinellen Lernens entwickelt, die für Datensätze eingesetzt werden können, die geschützte Merkmale enthalten. Dabei ergibt sich oft ein Zielkonflikt zwischen Fairness bzw. Diskriminierungsvermeidung (gemessen anhand von Fairnesskriterien, s. u.) einerseits und der Genauigkeit andererseits. Verschiedene diskriminierungsvermindernde Algorithmen schneiden dabei unterschiedlich gut hinsichtlich Fairness oder Genauigkeit ab (Übersicht und Test in Friedler u. a. 2019).

Ebenso werden Metriken entwickelt, mit denen die Fairness gemessen werden kann (**Fairnesskriterien bzw. -maße oder -metriken**). Nach Berk u. a. können die folgenden Fairnesskriterien unterschieden werden: (1) gleiche Genauigkeit („overall accuracity equality"), (2) statistische Parität („statistical parity"), (3) gleiche bedingte Gruppengenauigkeit („conditional procedure accuracy equality"), (4) gleiche bedingte Vorhersagegenauigkeit („conditional use accuracy equality") und (5) gleiches Fehlerverhältnis („treatment equality") (Berk 2018; Schweighofer u. a. 2018: 39f.).[97] Chouldechova weist am Beispiel der Algorithmen für die Berechnung von Rückfallquoten im Justizsystem (Beispiel 35, S. 66) jedoch darauf hin, dass die Fairnesskriterien nicht gleichzeitig erfüllt werden können (Chouldechova 2017). Welche Fair-

[95] Siehe auch Abschnitt 5.1.2, S. 79.
[96] Siehe auch Friedler u. a. (2019) oder Castelluccia & Le Métayer (2019: 46–47).
[97] Eine weitere Übersicht wird in Verma & Rubin (2018) gegeben.

nesskriterien in welchen Situationen zu welchen Differenzierungszwecken genutzt werden sollten, kann in der Informatik oder durch die Anwendenden nicht entschieden werden, sondern bedarf politischer Behandlung und Entscheidungen (Berk u.a. 2018; Castelluccia & Le Métayer 2019: 55). Zudem wird bei Betrachtung von Fairnesskriterien nicht die Anwendung des Systems selbst hinterfragt, sondern als gegeben vorausgesetzt. Gesellschaftliche Diskriminierungsrisiken, die durch die Anwendung von Differenzierungssystemen an sich entstehen, werden auf diese Weise nicht gelöst.

6.1.2 Verbesserungen des Nachweises von Diskriminierungen

6.1.2.1 Empirische Untersuchungen und Nachweise

Empirische Untersuchungen und Statistiken spielen seit langem eine bedeutende Rolle bei der Entdeckung und dem Nachweis von Diskriminierungen (z. B. Supik 2017). Sie stehen auch bei Entscheidungen von algorithmenbasierten Differenzierungen zur Erfassung und Auswertung der Konsequenzen und Ergebnisse zur Verfügung. Mit zunehmender Digitalisierung von verwaltungstechnischen und privatwirtschaftlichen Interaktionen liegen potenziell auch viel mehr Daten für statistische Analysen zur Identifizierung und dem Nachweis von Diskriminierungen vor.

Romei und Ruggieri (2014) liefern eine umfassende bibliografische Übersicht zu den gängigsten empirischen Diskriminierungsanalysen mit Verweisen auf exemplarische Studien. Dabei unterscheiden sie die Studien nach den Möglichkeiten, die die Forschenden haben, die Einflussvariablen (bzw. unabhängige Variablen) bei den statistischen Analysen zu beeinflussen. (1) Bei Beobachtungsstudien („observational studies“) haben die Forschenden keine Kontrolle über die Einflussvariablen. Sie erfassen die Daten von Beobachtungen durch Erhebungen oder Befragungen zu bestimmten Situationen, Zuständen, Strukturen von Wirtschafts- bzw. Lebensbereichen, wie z. B. Arbeits- oder Kreditmärkte, oder die Behandlungen von diskriminierungsgeneigten Personengruppen. Bei (2) quasi-experimentellen Studien („quasi-experimental studies“) haben Forschende nur über einige Einflussvariablen die Kontrolle. Zu dieser Studienart werden (2a) Auditstudien („auditing“) gezählt, bei denen Personen als Testpaare in Entscheidungssituationen geschickt werden und aus Vergleichen der Behandlung können Diskriminierungen abgeleitet werden. (2b) Bei Situationstests haben die Proband*innen Kontakt zur*m Entscheidungsträger*in und die und können (verborgene)

Aufzeichnungen über mögliche Ungleichbehandlungen durchführen. (2c) Korrespondenztests versuchen vor allem mit schriftlichen Anfragen, z. B. mit fingierten Bewerbungen und Lebensläufen, diskriminierendes Verhalten in den Reaktionen zu ermitteln. Vor allem dieser Studientyp wird im Onlinebereich eingesetzt, z. B. zur Untersuchung von Onlineangeboten der Personalsuche. (3) Bei experimentellen Studien haben die Forschenden die Kontrolle über alle Einflussvariablen. Hierbei können Laborexperimente und „realweltliche" Experimente unterschieden werden (Romei & Ruggieri 2014: 591–621). Sie verweisen zudem darauf, dass auch das Data-Mining zur Aufdeckung von Diskriminierungen geeignet ist (Romei & Ruggieri 2014: 621–624) etwa z. B. zur Aufdeckung von Diskriminierung nach Geschlecht bei Forschungsanträgen (Romei, Ruggieri & Turini 2013).

6.1.2.2 Algorithmen-Audits

Unter Algorithmen-Audits („algorithm audits") werden Methoden und Werkzeuge verstanden, die Forschende und Schutzeinrichtungen befähigen sollen, Systeme mit algorithmen- und datenbasierten Differenzierungen zu untersuchen und die helfen sollen, die Auswirkungen von Algorithmen, einschließlich Diskriminierungen, auf alle Typen von Betroffenen zu verstehen (Sandvig u. a. 2014; Hannák u. a. 2017: 2; Schweighofer u. a. 2018: 64–73). Teilweise entsprechen sie den Untersuchungsmethoden des Testens von Softwaresystemen oder den „klassischen" empirischen Untersuchungen von Diskriminierungen (s. o.). Nach der Einteilung von Sandvig (2014), die sich auf Onlineplattformen bezieht, gibt es (1) Audits des Programmcodes („code audits"), die dem Testen des Programmcodes mit der vollständigen Einsichtnahme entsprechen.[98] (2) Eine weitere Form erhebt die Daten über die Interaktionen von Plattformnutzenden und muss dazu keinen Einblick in den Code des Systems haben („noninvasive user audits"). (3) Des Weiteren werden Daten mit wiederholten Anfragen an eine Plattform gesammelt („scraping audits"). (4) Ebenso können Daten durch Nutzungen des untersuchten Dienstes durch fingierte Testpersonen, die von Computerprogrammen erzeugt werden und die die wiederholten Nutzungen durchführen, erhoben werden („sock puppet audit"). (5) Schließlich können relevante Daten mit der Nutzung des Dienstes durch Testpersonen, die über Crowdsourcingdienste angeheuert werden, erzeugt werden („crowdsourced audit" bzw. „collaborative audit").

[98] Zum Vergleich siehe Schweighofer u. a. (2018: 70).

Bei einigen der in Kapitel 4 beschriebenen Beispielfälle wurden Algorithmen-Audits zum Nachweis von Ungleichbehandlungen auf Webseiten, Onlineplattformen und Onlinemarktplätzen, wie z. B. für Onlinearbeitsmärkte (Beispiel 2, S. 35), bei Preisdifferenzierungen im Handel (Beispiel 12, S. 35) (siehe allgemein auch Mikians u. a. 2012, 2013) oder bei online verfügbaren Diensten der Gesichtserkennung (Beispiel 46, S. 74) eingesetzt.

Das Vorgehen in den oben genannten Fallbeispielen zur Aufdeckung von Ungleichbehandlungen und Diskriminierungen mithilfe von Algorithmen-Audits hat den Charakter von wissenschaftlichen Untersuchungen, die Expertise und vor allem Ressourcen für die empirische Erhebung von Interaktionen im Onlinebereich (z. B. Einsatz von Crawlern, Umgang mit fingierten Konten) und vor allem der statistischen Auswertung erfordern. Sie nutzen meist im Internet beschaffbare Informationen. Dazu werden die notwendigen Daten teils mit Webcrawlern zusammengetragen, teils durch umfangreiche automatisierte Anfragen an Suchmaschinen oder Onlineplattformen mithilfe von Software oder durch „Crowdworker" (z. B. über Amazon Mechanical Turk) erarbeitet.[99] Es wird auch sichtbar, dass eine wachsende Anzahl von Forschenden die Identifikation und Untersuchungen von Diskriminierungen, vor allem im Onlinebereich, zum Gegenstand ihrer Forschungen machen und den Methodenbaukasten dazu weiterentwickeln.

Die Beispiele 16 (S. 46) mit dem System „Sunlight" und 17 (S. 46) mit dem System „AdFisher" liefern auch Werkzeuge der automatisierten Datenerfassung und -auswertung von Online-Transaktionen (teils mit maschinellen Lernverfahren), die der Untersuchung von Personalisierungen oder potenziell diskriminierenden Differenzierungen dienen können. Derartige Werkzeuge können der Durchführung von Algorithmen-Audits (insbesondere Typ 2 und 3) dienen.

Schlussfolgerungen

- Forschungsergebnisse und Beispielfälle zeigen, dass der Nachweis von Ungleichbehandlungen und Diskriminierungen, die auf algorithmen- und datenbasierten Differenzierungen beruhen, auch ohne den Zugriff und die direkte Inspektion der Algorithmen

[99] Siehe z. B. Beispiel 2 (S. 35), 8 (S. 40), 12 (S. 44) oder 38 (S. 69).

erfolgen kann. Dies wird durch die Beobachtung, Erfassung und Auswertung der Ergebnisse und Konsequenzen der Differenzierungsanwendungen ermöglicht, häufig, indem die Rolle als Testnutzende eingenommen wird, auch kann dies über automatisierte Nutzungen und Abfragen erfolgen.

- Ebenso zeigen Beispiele[100], dass es gerade bei Systemen der automatisierten Entscheidungsunterstützung bedeutend sein kann, das Gesamtergebnis, d. h., welche tatsächlichen Konsequenzen durch menschliche Entscheidungen angesichts der Computerempfehlungen ausgelöst werden, in Augenschein zu nehmen. Dies ist auch für die vielen Beispiele relevant, bei denen KI-Systeme bestimmte Personengruppen mit geschützten Merkmalen schlechter erkennen (in Abschnitt 4.12, S. 68). Bei ihnen würden Diskriminierungen erst deutlich, wenn bestimmte auf der schlechteren Erkennung basierende Praktiken und Entscheidungen betrachtet würden (und diese z. B. zu überproportional mehr Polizei- oder Grenzkontrollen, unterproportional weniger Personaleinstellungen oder unangemessener Berücksichtigung in Marketingstrategien führen würden). Jedoch setzt auch die Form der Nachweise über die Ermittlung von Ergebnissen Fachkenntnisse voraus, vor allem in der Statistik und Programmierung, ebenso wie finanzielle und personelle Ressourcen für die Durchführung der Untersuchungen. Daher dürfte diese Nachweisform für betroffene Individuen ohne Fachkenntnisse nicht geeignet sein. Für Antidiskriminierungsstellen, insbesondere in Kooperation mit Forschungseinrichtungen, können die Nachweisformen geeignete Instrumente sein und werden von ihnen auch in Anfängen genutzt.[101]

- Eine weitere Voraussetzung ist, dass die relevante Kommunikation und die Transaktionen zu den Differenzierungen auch statistisch erfassbar sind, wie dies für viele Onlineanwendungen und Onlineplattformen, die öffentliche Angebote haben, möglich scheint. Derartige Nachweismöglichkeiten enden jedoch dort, wo eben diese Transaktionen und Kommunikation nicht öffentlich erfassbar sind, wie bei geschlossenen Verwaltungsvorgängen oder bei exklusiv-individualisierten kommerziellen Angeboten.

---

[100] Siehe Beispiele zu den Arbeitsvermittlung in Polen (Beispiel 29, S. 57) und Österreich (Beispiel 30, S. 59).

[101] Siehe z. B. Fallbeispiel 47, S. 75.

Des Weiteren können auf diesen Wegen nicht die genauen Ursachen von Diskriminierungen aufgedeckt werden, wenn mehrere Ursachen in einem komplexen System oder im komplexen Wechselspiel zwischen Datenssätzen, Algorithmen und menschlichen Entscheidenden liegen können.[102] Dazu bieten sich die Möglichkeiten der Untersuchung von verwendeten Praktiken der Datenerzeugung (wie etwa im Beispiel 34) und den Möglichkeiten des Testens von Softwaresystemen und der Simulation von Variationen der Inputdaten oder der Komponenten von Entscheidungsregeln an, genauso wie Untersuchungen dazu, wie menschliche Entscheidende mit Computerempfehlungen umgehen.

6.1.3 Rechtliche Situation

6.1.3.1 Informationspflichten und Auskunftsrechte des Datenschutzes

Das Datenschutzrecht umfasst verschiedene Informationspflichten für die verantwortlichen Stellen bzw. Betreibenden von Datenverarbeitungen gegenüber den Betroffenen (Art. 12, 13 und 14 DSGVO) sowie Auskunftsrechte, die die Betroffenen gegenüber den Betreibenden geltend machen können (Art. 15 DSGVO). Die Informationspflichten dienen dazu, dass die Betroffenen von der Datenverarbeitung erfahren, damit sie ihre Rechte effektiv wahrnehmen können. Ebenso sollen sie die Basis der Einwilligung durch die Betroffenen bilden, mit der (neben anderen Gründen) die Rechtmäßigkeit der Datenverarbeitung bestimmt wird (Art. 6 Abs. 1 DSGVO). Daher wird auch von **„informierter Einwilligung“** gesprochen. Mit den Auskunftsrechten haben die Betroffenen das Recht, Auskunft über den Zweck und die Reichweite der Datenverarbeitung zu erlangen, damit gewährleistet ist, dass sie prüfen können, ob die Daten rechtmäßig verarbeitet wurden (Busch 2018: 37–41). Beim Vorliegen von automatisierten Entscheidungen gelten erweiterte Informationspflichten (siehe Abschnitt 6.2.3, S. 114ff.).

Die Konkretisierung der Informationspflichten und des Konzepts der informierten Einwilligung in der Praxis erfolgt dabei in der Regel mithilfe der Datenschutzerklärung oder Allgemeinen Geschäftsbedingungen (AGB),

[102] Dies zeigt z. B. das Fallbeispiel 17, S. 46.

deren Ausgestaltung jedoch kritisiert wird. Dabei werden die Uneindeutigkeiten der verwendeten Begriffe hervorgehoben, die Verwendung von nicht leicht zu verstehender juristischer Sprache, die Unzulänglichkeiten der gelieferten Informationen, kognitive Hürden, diese zu verstehen und Aufwand und Zeitbeschränkungen, die einem Lesen und Verstehen der zahlreichen Datenschutzerklärungen entgegenstehen (Milne & Culnan 2004; Solove 2013; Cate & Mayer-Schönberger 2013; Reidenberg u.a. 2015; Reidenberg u.a. 2016; McDonald & Cranor 2008; Van Alsenoy, Kosta & Dumortier 2014; Martin 2013; Moll u.a. 2018; Kamp & Rost 2013; Orwat & Schankin 2018; Kettner, Thorun & Kleinhans 2018; Hänold 2019).

Zwar muss in der Datenschutzerklärung der Zweck der Datenverarbeitung[103] genannt werden, aber es darf bezweifelt werden, dass aus diesen Angaben die Konsequenzen der auf der Datenverarbeitung beruhenden Differenzierungsentscheidungen im Sinne einer Ungleichbehandlung durch die Betroffenen abgeschätzt werden können. Da es hier nur um die Informierung über die Existenz eines beabsichtigten Verarbeitungszwecks geht und nicht über dessen Konsequenzen, kann man vermuten, dass sich mit diesem Rechtsinstrument nicht die Indizien finden lassen, die für ein Antidiskriminierungsvorgehen notwendig sind.

6.1.3.2 Beweislast und Indizien nach dem AGG

Nach § 22 AGG ist eine Beweiserleichterung für die Betroffenen vorgesehen, 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 an drei Voraussetzungen geknüpft: (1) Die Person, die behauptet, diskriminiert worden zu sein, muss nachweisen, dass sie anders behandelt wurde als andere Personen, und (2) sie muss nachweisen, dass sie sich im Hinblick auf eines der geschützten Merkmale (nach § 1 AGG) unterscheidet. (3) Ferner müssen von der Person Indizien erbracht werden, die mit überwiegender Wahrscheinlichkeit darauf schließen lassen, dass das in § 1 genannte Merkmal ursächlich für die Diskriminierung war (Ebert 2019: § 22 AGG Rn. 1 und 2). Diese Anforderungen sind bezüglich algorithmischer Diskriminierungsrisiken **problematisch**:

[103] Nach Art. 13 Abs. 1 lit. c) DSGVO.

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. Bei Personalisierung und dem zielgerichteten, exklusiven Angebot von Informationen, Diensten oder Produkten, insbesondere in Onlineangeboten und auf Onlineplattformen, kann eine einzelne, möglicherweise betroffene Person den Nachweis schlecht erbringen, dass sie gegenüber Vergleichspersonen schlechter behandelt wird. Wenn sich die Angebote zudem noch dynamisch ändern, wird der Nachweis weiter erschwert. Noch gravierender ist, dass der Nachweis einer Ungleichbehandlung, die im Hinblick auf das geschützte Merkmal erfolgt, eben durch die unabsichtliche oder absichtliche, verschleiernde Nutzung von Ersatzvariablen bzw. Proxies, die mit dem geschützten Merkmal korrelieren, für eine Einzelperson ohne tiefe Fachkenntnis nicht möglich ist (Abschnitt 5.1, S. 77ff.). So kann das Individuum auch kaum die notwendigen Indizien erbringen. Dies zeigen etwa die aufwendigen empirischen Untersuchungen und Algorithmen-Audits bei den Beispielfällen, die notwendig sind, damit überhaupt tendenziell Diskriminierungen nachgewiesen werden konnten (Kapitel 4).[104] Neben diesen Problemen können (antizipierte) Abhängigkeitsverhältnisse (wie z. B. bei der Auswahl der Bewerbenden) oder das hohe Prozesskostenrisiko Hindernisse sein, gegen Diskriminierungen vorzugehen, falls sie für die Betroffenen wahrnehmbar sein sollten.

Schlussfolgerung

Eine Lösungsmöglichkeit bestünde im **kollektiven Rechtsschutz** mit der Verbandsklage, die für das Antidiskriminierungsvorgehen seit längerem gefordert wird (Berghahn u. a. 2016: 141ff., 159–162; Ponti & Tuchtfeld 2018; Straker & Niehoff 2018). Jedoch bleibt die Notwendigkeit bestehen, dass eine potenzielle Diskriminierung erst einmal durch irgendjemand als mögliche Schädigung wahrgenommen werden muss, von dem dann die Initiative zur Verbandsklage ausgehen kann. Für die schwer wahrnehmbaren algorithmischen Diskriminierungsrisiken wären (ergänzende) regelmäßige Untersuchungen besser geeignet, die durch Antidiskriminierungs-, Forschungs- oder ähnliche Einrichtungen oder spezialisierte Behörden (vor allem auch in Kooperationen) durchgeführt werden, ohne

[104] Siehe ähnlich auch Hänold (2019) mit Bezug auf Algorithmen, Profiling und Scoring im Versicherungswesen. Siehe auch Fröhlich & Spiecker genannt Döhmann (2018).

dass ein konkreter Schaden als Anlass vorhanden sein muss. Zudem können rechtlich gesicherte Auskunftsansprüche für die Antidiskriminierungsstellen ein solches Vorgehen zur Erbringung der Beweise erleichtern, die allerdings bisher noch nicht vorhanden sind.

6.1.3.3 Dokumentationen

Aus der Umkehr der Beweislast folgern Dzida und Groh für den Arbeitsbereich, dass im Fall des Rechtsstreits, bei dem ein Gericht eine Benachteiligung vermutet, auch die Anwendenden von Algorithmen Schwierigkeiten haben, zu beweisen, dass keine Diskriminierung vorliegt, oder dass die Verhältnismäßigkeit des Einsatzes des Systems für die Differenzierungsaufgabe, wie z. B. die Personalauswahl, gegeben gegeben sei (Dzida & Groh 2018). So kann bei der Darlegung der Verhältnismäßigkeit möglicherweise zwar die Eignung des Systems für die Differenzierungsaufgabe nachgewiesen werden, etwa durch wissenschaftliche Untersuchungen, aber hinsichtlich des Nachweises der Erforderlichkeit und Angemessenheit darf das jeweilige Differenzierungsziel „[...] nicht mit anderen gleichermaßen geeigneten, aber weniger einschneidenden Mitteln zu erreichen sein und die Verfahren dürfen keine übermäßige Beeinträchtigung legitimer Interessen benachteiligter Personen mit sich bringen." (ebd., S. 1920f.).

Für die Beweispflicht, dass keine Diskriminierungen vorliegen, kann es für die Anwendenden notwendig werden, Vorkehrungen der Nachvollziehbarkeit von Algorithmen bzw. künstlicher Intelligenz zu treffen, damit in Fällen des Rechtsstreits nachgewiesen werden kann, wie eine Entscheidung und deren Ergebnisse bzw. Konsequenzen für Betroffene zustande gekommen sind, oder nach welchen Entscheidungskriterien und Gewichtungen Differenzierungsentscheidungen getroffen wurden. Nach Yeung können sich diese Anforderungen auch aus dem Prinzip der Verfahrensgerechtigkeit ergeben, nach dem Personen bei Gerichtsverfahren ein Recht haben, die Gründe für Entscheidungen, die sie nachteilig und erheblich betreffen, zu erfahren (Yeung 2017: 23)[105].

[105] Siehe dazu auch die Gerichtsverfahren in den USA; siehe AI Now Institute (2018).

So werden Protokollpflichten über die Programmabläufe oder verwendeten Merkmale der Differenzierungsentscheidungen vorgeschlagen, die zum Nachweis in Streitfällen dienen sollen (Martini 2017; Ernst 2017: 1032; Brauneis & Goodman 2018). Des Weiteren könnten die Regelungen zum Verfahrensverzeichnis (nach Art. 32 DSGVO) sowie zu den Verhaltensregeln und zur Zertifizierung (Art. 40 DSGVO) dahingehend weiterentwickelt werden, dass sie die Besonderheiten algorithmischer Entscheidungsverfahren berücksichtigen.

6.2 Detailliertere Regulierung von algorithmischen Entscheidungsregeln

Angesichts der Probleme der Informationspflichten und Auskunftsrechte, die keinen ausreichenden Selbstschutz vor Benachteiligungen und Diskriminierungen ermöglichen, sowie angesichts des schwierigen, wenn nicht sogar unmöglichen Nachweises von Diskriminierungen durch Betroffene (ohne Expertise), sollte über eine stärkere Regulierung der Differenzierungsentscheidungen nachgedacht werden.[106] Regulierungen von Differenzierungsentscheidungen, die auf den Auswertungen personenbezogener Daten beruhen, sind nicht neu. Sie finden sich in dem für die Antidiskriminierung und für die informationelle Selbstbestimmung relevanten Recht und dessen institutioneller Umsetzung in einschlägigen Behörden und Einrichtungen. So kann beispielsweise auch das AGG als eine Form der Regulierung von Entscheidungsregeln verstanden werden, bei der die Verwendung bestimmter Entscheidungsmerkmale ausgeschlossen ist.

6.2.1 Benachteiligungsverbote und geschützte Merkmale

Mit dem Allgemeinen Gleichbehandlungsgesetz (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: 4). So ist grundsätzlich die unmittelbare

[106] Siehe auch Raabe & Wagner (2019 in Vorbereitung).

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: 1032).

Die bestehenden Kataloge von geschützten Merkmalen des AGG[107] sollten dahingehend geprüft werden, ob sie die Merkmale abdecken, die mit (neuen) algorithmischen Verfahren identifiziert werden können, und ob neue Diskriminierungsgründe geschaffen werden. Denn Systeme mit maschinellem Lernen und anderen Formen der KI ermöglichen es, Merkmale zu identifizieren und nach ihnen zu differenzieren, die bisher nicht in den Katalogen des Antidiskriminierungsrechts enthalten sind (vgl. auch Zuiderveen Borgesius 2018: 20). Die Systeme können Analysemöglichkeiten zum Erkennen von Stimmungen, der Naivität und Beeinflussbarkeit, Erkennen von kognitiven Schwächen oder psychischen und emotionalen Zuständen (wie z. B. Depressionen), den jeweiligen sozialen Status oder Charaktereigenschaften umfassen (siehe Abschnitt 2.2.2, S. 7). Dabei kann es sich beispielsweise um die Merkmale „biometrische Merkmale“, „politische Meinung“ oder „Gesundheitszustand“ handeln, die zwar in der DSGVO als besondere Kategorie personenbezogener Daten (s. u.) geregelt sind, aber nicht im AGG.

Es kann im Moment über die gesellschaftlichen Folgen nur gemutmaßt werden, da Erkenntnisse über den tatsächlichen Einsatz der Systeme und die ausgelösten Veränderungen noch weitgehend fehlen. Hierzu sind Forschungen notwendig, die potenzielle Anwendungen derartiger Systeme, Wirkmechanismen und Gefährdungspotenziale auf Gleichbehandlung und freie Persönlichkeitsentfaltung untersuchen. Jedoch dürfte schon jetzt ein Risiko, dass derartige identifizierte Persönlichkeitsmerkmale ausgenutzt werden könnten, um Personen nach ihrer Angewiesenheit auf ein Gut, eine Ressource oder Position (z. B. Arbeitsstelle) zu differenzieren und damit die strukturelle Überlegenheit[108] der Anbietenden zu erhöhen, nur

[107] Zur Diskussion der Merkmalskataloge, siehe z. B. Däubler (2018: AGG § 1 Rn. 6–10).
[108] Siehe Abschnitt 5.2.6, ab S. 95.

schwer von der Hand zu weisen sein. Das Antidiskriminierungsrecht könnte mit der Ausweitung dem Katalog der geschützten Merkmale zum Schutz vor Missbrauch der strukturellen Überlegenheit genutzt werden.

Ferner ist in diesem Zusammenhang zu prüfen, ob die Regelungen des Art. 9 DSGVO ausreichend für den Schutz gegen algorithmische Diskriminierungen sind. Danach ist die **Verarbeitung sensibler personenbezogener Daten**, „[...] aus denen die rassische und ethnische Herkunft, politische Meinungen, religiöse oder weltanschauliche Überzeugungen oder die Gewerkschaftszugehörigkeit hervorgehen, sowie die Verarbeitung von genetischen Daten, biometrischen Daten zur eindeutigen Identifizierung einer natürlichen Person, Gesundheitsdaten oder Daten zum Sexualleben oder der sexuellen Orientierung“ (Art. 9. Abs. 1 DSGVO) untersagt, es sei denn, einer der zahlreichen Ausnahmen gestattet es. So ist die Verarbeitung von sensiblen Daten zulässig, wenn die betroffene Person ausdrücklich einwilligt (Art. 9 Abs. 2 DSGVO). Auch hierbei müssten durch die betroffene Person prinzipiell die teils langfristigen individuellen Konsequenzen, auch im Sinne einer Ungleichbehandlung, im Zeitpunkt der Einwilligungsabgabe abgeschätzt werden, was jedoch eine große Herausforderung sein kann. Ob daraus eine realistische Chance auf Selbstschutz der Betroffenen vor algorithmischen Diskriminierungen erwächst, kann bezweifelt werden.

Zusammenfassung und Schlussfolgerung

Die im GG und AGG festgelegten Kataloge geschützter Merkmale sind dahingehend zu prüfen, ob neue Analysemethoden, vor allem mit Algorithmen der künstlichen Intelligenz, zur automatisierten Identifizierung von Persönlichkeitsmerkmalen ihre Erweiterung erfordern. Dabei werden auch Merkmale identifizierbar und für Differenzierungen zugänglich, die dazu genutzt werden können, die Angewiesenheit auf ein Gut, eine Ressource oder eine Position zu ermitteln und zu nutzen, um strukturelle Überlegenheit aufzubauen oder zu erhöhen. Die noch weitgehend unbekannten Zusammenhänge zwischen technisch machbaren und potenziell gefährdeten Schutzzielen sollten erforscht und ihre Legitimität sollte gesellschaftlich beurteilt werden.

6.2.2 Ausnahmen nach sachlichem Grund und anerkannten Methoden

Nach § 20 Abs. 1 AGG liegt kein Verstoß des Benachteiligungsverbots vor, wenn ein sachlicher Grund vorliegt. Der unkonkrete Begriff „sachlicher Grund" wird von den in den nachfolgenden Sätzen angeführten Beispielen in seiner Dimension verdeutlicht, aber nicht abschließend geklärt (Schrader & Schubert 2018: AGG § 3 Rn. 68ff.). Danach liegt ein sachlicher Grund beispielsweise dann vor, wenn (a) es um die Vermeidung von Gefahren, Verhütung von Schäden oder Ähnlichem geht, wenn z. B. aufgrund von Verkehrssicherungspflichten bestimmte Personengruppen von der Benutzung bestimmter Geräte oder Fahrzeuge ausgeschlossen werden. Hier ist es allerdings strittig, inwieweit auch eine wirtschaftliche Gefährdung, z. B. in Form von Umsatzeinbußen, dazu zählen.[109] (b) Ebenso kann ein sachlicher Grund vorliegen, wenn bestimmte Personen zum Schutz der Intimsphäre oder der persönlichen Sicherheit anderer Personen ausgeschlossen werden (z. B. bei nach Geschlechtern getrennten Öffnungszeiten von Schwimmbädern oder Saunen). (c) Ein weiteres Beispiel für einen sachlichen Grund sind besondere Vorteile bzw. Vergünstigen und wenn bei ihnen kein Interesse an der Durchsetzung der Gleichbehandlung besteht, wie z. B. bei aus sozialen Gründen gewährten Preisnachlässen (z. B. für Studierende) oder bei begünstigenden Verkaufsfördermaßnahmen, die nur bestimmte Personengruppen betreffen (z. B. günstigere Preise für Männer bei Tanzkursen, bei denen ein Frauenüberhang besteht, oder umgekehrt) (Franke & Schlichtmann 2018: AGG § 20 Rn. 12–21). Liegen diese sachlichen Gründe nicht vor, bleibt nur die Klärung im Einzelfall, wobei Abwägungen nach Verhältnismäßigkeitsgrundsätzen getroffen werden müssen (Schrader & Schubert 2018: AGG § 3 Rn. 68ff.). Die Rechtssituation liefert jedoch Unsicherheiten über die Auslegung im Vorhinein, etwa, wenn es bei der Gestaltung von Entscheidungssystemen darum geht, welche Merkmale verwendet werden können oder nicht.

§ 20 Abs. 2 Satz 2 AGG regelt die Ungleichbehandlung bei Versicherungsverträgen hinsichtlich der Merkmale Religion, Behinderung, Alter und sexuelle Identität (eine Differenzierung nach allen anderen geschützten Merkmalen ist nach § 19 AGG verboten). Sie ist dann zulässig, wenn sie nach „anerkannten Prinzipien risikoadäquater Kalkulation" erfolgt, insbesondere wenn sie auf „[…] einer versicherungsmathematisch ermittelten Risikobewertung

[109] So sehen dies Franke & Schlichtmann (2018: AGG §20 Rn. 17) sehr restriktiv.

unter Heranziehung statistischer Erhebungen“ (§ 20 Abs. 2 Satz 2 AGG) basiert. Schiek (2000: § 20 AGG Rn. 8) hält die Versicherungsdiskriminierung mit der fortgesetzten Bildung von Stereotypen und Vorurteilen durch die Erhebung von Statistiken in Anknüpfung an geschützte Merkmale für eine sachlich nicht gerechtfertigte Durchbrechung des Gleichbehandlungsanspruchs. Eine erweiterte Anwendung dieser Form der Legitimation auf andere Lebensbereiche, wie z. B. Bankdienstleistungen, hält sie für unzulässig. Auch Berghahn u. a. fordern die Einschränkung dieser Form der Ungleichbehandlung (Berghahn u. a. 2016: 122ff.). Dennoch werden in der Praxis des Scorings z. B. im Onlinehandel geschützte Merkmale mit Verweis auf ökonomische Interessen als sachlicher Grund benutzt (z. B. Moos & Rothkegel 2016).

Schlussfolgerungen

Mit der zunehmenden Verbreitung von algorithmischen Berechnungen und Durchsetzungen von Differenzierungen, einschließlich Scoring, sollte geprüft werden, ob die Zulässigkeiten der Berechnungsmethoden, die Begründungen, die Merkmale, die Anwendungsfelder und Differenzierungszwecke nicht eindeutiger und verbindlicher geregelt werden sollten. Das sollte auch die Verfahren der Anerkennung bei den „anerkannten Prinzipien risikoadäquater Kalkulation“ bzw. „wissenschaftlich anerkannten mathematisch-statistischem Verfahren“ betreffen. Angesichts der rapide steigenden Anzahl von Algorithmen, maschinellen Lernverfahren, Konkretisierungen einer „hinreichenden“ Prognosegenauigkeit, Fairnesskriterien und Qualitäts- bzw. Fehlermaßen ist es wahrscheinlich, dass auch die Einschätzungen über ihre jeweilige Tauglichkeit und Adäquatheit auseinanderlaufen können, sodass man nicht mehr von einer allgemein anerkannten Methode sprechen kann. Eindeutigere und allgemein verbindliche Klärungen dazu könnten stabile Erwartungen bei den Entwickelnden, Anwendenden und gegebenenfalls Betroffenen erzeugen.

6.2.3 Verbot automatisierter Entscheidungen

Eine der wichtigsten Regelungen zum Schutz vor algorithmischen Diskriminierungsrisiken ist das Verbot von automatisierten Entscheidungen im Datenschutzrecht. Der Zweck der Vorschrift findet sich bereits in der Datenschutz-Richtlinie (DSRL, RL 95/46/EG) und dem alten Bundesdaten-

schutzgesetz (BDSG a. F.) und dient dort dem Schutz der menschlichen Individualität als Element des Rechts auf freie Entfaltung der Persönlichkeit und autonomen Gestaltung des eigenen Lebens (Ernst 2017: 1030; Martini 2018: DSGVO Art. 22 Rn. 8; Hoeren & Niehoff 2018: 53). Nach Art. 22 Abs. 1 DS-GVO 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 DSGVO geht nicht eindeutig hervor, welche Arten von automatisierten Entscheidungen tatsächlich erfasst sind. Das kann nur indirekt aus dem Wortlaut der Rechtsnorm abgeleitete werden:

(1) Zum einen sind dies die automatisierten 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 (Ernst 2017: 1029f., 1031; Busch 2018: 31). Martini betont, dass es maßgeblich sei, ob ein Mensch auf die Entscheidung, d. h. deren Inhalt, Einfluss nimmt. Dabei kann der Mensch zwar Entscheidungen manuell vorbereiten, denn ob das Verbot gilt, hängt nicht von der Vorbereitung, sondern von der Entscheidung selbst ab. Hat ein Mensch eine inhaltliche Entscheidungsbefugnis, übt er diese Entscheidungsmacht tatsächlich aus, und kommt es zu einem regelmäßigen Eingreifen, d. h., es liegen keine Strichprobenkontrolle und zudem kein Einzelfall des menschlichen Eingreifens vor, kann nicht mehr von einer ausschließlich automatisierten Entscheidung gesprochen werden und das Verbot gilt nicht (Martini 2018: DS-GVO Art. 22 Rn. 17–19).[110]

(2) Zum anderen sind alle die Arten von automatisierten Entscheidungen erfasst, die „rechtliche Wirkung entfalten oder die Person in ähnlicher Weise erheblich beeinträchtigen" (Scholz 2019: DSGVO Art. 22 Rn. 31–37). Be-

[110] Ähnlich auch Weichert (2018: 128–135, insbesondere S. 133f.) für den als verbotene automatisierte Entscheidungen auch Situationen gelten, in denen die natürliche Person Unterlagen vor der Entscheidung nur prüft oder rein formal in den Entscheidungsprozess involviert ist. Ähnlich äußern sich Hoeren & Niehoff (2018: 53), ebenso Scholz: „Von einer ausschließlich automatisierten Entscheidung ist nicht nur dann auszugehen, wenn von vornherein keine Überprüfung durch einen Menschen vorgesehen ist und eine solche nicht stattfindet, sondern auch, wenn der Mensch – ohne eigene Erwägungen anzustellen – die automatisierte Vorgabe lediglich bestätigt oder übernimmt." Scholz (2019: DSGVO Art. 22 Rn. 26).

stimmend ist demnach die Art der Wirkungen.[111] Die „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 gegebenen, wenn die betroffene Person in ihrer wirtschaftlichen und persönlichen Entfaltung erheblich gestört wird, wie z. B. bei Versagen eines günstigen Zinssatzes (Busch 2018: 33).

Liegen automatisierte Entscheidungen vor, die nach den genannten Artikeln der DSGVO erlaubt sind, dann sind weitere Regelungen einzuhalten (Weichert 2018: 131; Hoeren & Niehoff 2018: 54f.): Art. 14 Abs. 2 lit. g DSGVO regelt die **Informationspflichten**, dass beim Bestehen 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 (Weiteres siehe unten). Die **Auskunftsrechte** der betroffenen Person werden in Art. 15 Abs. 1 lit. h DSGVO geregelt und sehen, mit gleichem Wortlaut, Information über die involvierte Logik und Auswirkungen vor.

Zudem wird bei automatisierten Entscheidungen eine **Datenschutz-Folgenabschätzung**[112] nach Art. 35 Abs. 3 lit. a DSGVO 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-Folgenab-

[111] Eine andere Interpretation liefert Weichert (2018) mit Verweis auf Buchner (2018: Art. 22 Rn. 18) und Reichwald & Pfisterer (2016: 211f.), der das Verbot insbesondere aus dem Grad der Intransparenz und fehlenden Beeinflussbarkeit für die Betroffenen her bestimmt sieht. So fallen nur automatisierte Entscheidungssysteme, bei denen der Entscheidungsprozess für die Betroffenen nicht mehr überschaubar ist und Kontrollierbarkeit und Revisionsfähigkeit für die Betroffenen fehlen, unter das Verbot. Das kann vorliegen, wenn die Algorithmen nicht vollständig dokumentiert werden oder bei automatisierten Entscheidungen mit lernenden Algorithmen bzw. künstlicher Intelligenz (Weichert 2018: 130). Allerdings können, wie bei den Diskriminierungsbeispielen gezeigt, auch vergleichsweise „einfache" Algorithmen, ohne Involvierung von Menschen und auf vertraglicher Basis, rechtlich nachteilige oder erheblich beeinträchtigende Wirkungen haben, die nach dieser Interpretation nicht unter das Verbot fallen würden.

[112] Genauere Vorgaben zur Datenschutz-Folgenabschätzung werden durch die Article 29 Data Protection Working Party, WP29 (2017b) gegeben.

schätzung diese Risiken vorab bewerten, ebenso, ob die Verarbeitungsvorgänge notwendig und verhältnismäßig sind, genauso wie die zur Bewältigung der Risiken geplanten Abhilfemaßnahmen (Art. 35 Abs. 7 lit. b bis d DSGVO). Das bedeutet beispielsweise, dass die verantwortliche Stelle auch den Einsatz von Verfahren des maschinellen Lernens bzw. der künstlichen Intelligenz selbst unterlässt (bzw. unterlassen muss), wenn das Risiko als unverhältnismäßig hoch zum Verarbeitungszweck abgeschätzt wird und weniger risikobelastete Alternativen des Entscheidungsverfahrens verfügbar sind bzw. das fragliche Verfahren nicht unbedingt für den (Differenzierungs-)Zweck notwendig ist. Werden hohe Risiken festgestellt, muss eine Meldung an die Aufsichtsbehörde erfolgen (nach Art. 36 DSGVO). Die Aufsichtsbehörde kann in diesen Fällen die Verarbeitung untersagen (nach Art. 58 Abs. 3 lit. f DSGVO).

6.2.3.1 Ausnahmen

Ausnahmen vom Verbot automatisierter Einzelentscheidungen werden in Abs. 2 geregelt. Sie liegen vor, wenn die automatisierte Entscheidung zum Abschluss oder zur Erfüllung eines Vertrages erforderlich ist, durch Rechtsvorschriften der Europäischen Union oder der Mitgliedstaaten oder durch **ausdrückliche Einwilligung**[113] der betroffenen Person zulässig ist. Nach Art. 22 Abs. 4 DSGVO gelten diese Ausnahmen jedoch nicht, wenn Entscheidungen auf der Verarbeitung von **besonderen Kategorien personenbezogener Daten** des Art. 9 DSGVO beruhen („sensible Daten"), was ausdrücklich der Antidiskriminierung dienen soll (Buchner 2018: DS-GVO Art. 22 Rn. 44). Dieses Verbot wird jedoch wieder eingeschränkt, wenn eine ausdrückliche Einwilligung der betroffenen Person vorliegt (Art. 9 Abs. 2 lit. a DSGVO), oder wenn die Verarbeitung aus Gründen eines erheblichen öffentlichen Interesses erforderlich ist (Art. 9 Abs. 2 lit. a DSGVO). Greift bei der Verarbeitung von besonders schützenswürdigen Daten eine dieser beiden Ausnah-

[113] Der Begriff „ausdrückliche Einwilligung" wird nicht in der DSGVO erläutert. Die Richtlinie der Article 29 Data Protection Working Party zur Einwilligung gibt Erläuterungen dazu; siehe WP29 (2017a: 18f.). Scholz führt dazu aus: „Auch wenn es hier formal gesehen nicht um die Einwilligung in einzelne Datenverarbeitungsschritte, sondern um die Anwendung eines Datenverarbeitungsverfahrens geht, wird man diese Einwilligung ebenfalls an den Voraussetzungen der Art. 4 Nr. 11 und Art. 7 messen müssen.[...] Aus der Perspektive der betroffenen Personen ist das Schutzbedürfnis vergleichbar. Die Einwilligung ist daher nur wirksam, wenn sie unmissverständlich, freiwillig, bestimmt sowie informiert abgegeben wird [...]. Letzteres setzt voraus, dass die betroffene Person noch vor der Einwilligung alle Informationen bekommen muss, die notwendig sind, um Anlass, Ziel und Folgen der Verarbeitung korrekt abschätzen zu können" Scholz (2019: DSGVO Art. 22 Rn. 52).

men, hängt die Zulässigkeit der automatisierten Entscheidung zusätzlich davon ab, ob die Ausnahmen des Art. 22 Abs. 2 gelten, also, ob sie für den Abschluss oder die Erfüllung eines Vertrages erforderlich ist oder ob eine ausdrückliche Einwilligung der betroffenen Person vorliegt (Buchner 2018: DS-GVO Art. 22 Rn. 45f.; Busch 2018: 35). In diesem Zusammenhang wird von Ernst auf eine Dilemmasituation hingewiesen: Zwar ist die Einwilligung zu einer Datenverarbeitung nach Art. 22 Abs. 2 lit. c DSGVO möglich, diese steht aber in Konflikt mit dem AGG, das eine Benachteiligung auch dann ausschließt, wenn eine Einwilligung zur Ungleichbehandlung vorliegt (Ernst 2017: 1033; Schrader & Schubert 2018: AGG § 3 Rn. 47).

6.2.3.2 Angemessene Maßnahmen

Auch soll bei den Ausnahmefällen des Art. 22 Abs. 2 lit. a und c DSGVO, also bei Vorliegen von „Vertrag“ und „Einwilligung“, d.h. den Situationen, in denen eine automatisierte Entscheidung erlaubt ist, die automatisierte Entscheidung mit angemessenen Maßnahmen durch die verantwortliche Stelle zur Wahrung der Rechte und Freiheiten sowie berechtigten Interessen der betroffenen Personen erfolgen. Dazu gehören zumindest das Recht der betroffenen Personen ein direktes Eingreifen einer Person der verantwortlichen Stelle zu erwirken, sowie das Recht auf Darlegung des eigenen Standpunktes und auf Anfechtung der Entscheidung (Art. 22 Abs. 3 DSGVO).[114] Das Ziel dieser Regelungen ist dabei nicht nur der Schutz vor diskriminierenden automatisierten Entscheidungen, sondern auch die Transparenz und Fairness bei der Entscheidungsfindung selbst (Scholz 2019: DSGVO Art. 22 Rn. 3, 56).

Was das Recht der betroffenen Person auf das **„direkte Eingreifen“** bzw. die sich daraus ergebende Möglichkeit, jederzeit der automatisierten Entscheidung zu widersprechen („opt out“), tatsächlich bedeutet, ist allerdings noch unklar. So legen etwa Martini und Nink sowie Busch dies so eng aus, dass nur bei berechtigten Gründen und im Einzelfall das Eingreifen einer Person verlangt werden kann (Martini & Nink 2017; Busch 2018: 36). Allerdings könnten die Absätze auch anders interpretiert werden. Denn das direkte

[114] Auch beschrieben im Erwägungsgrund 71 der DSGVO mit: „In jedem Fall sollte eine solche Verarbeitung mit angemessenen Garantien verbunden sein, einschließlich der spezifischen Unterrichtung der betroffenen Person und des Anspruchs auf direktes Eingreifen einer Person, auf Darlegung des eigenen Standpunktes, auf Erläuterung der nach einer entsprechenden Bewertung getroffenen Entscheidung sowie des Rechts auf Anfechtung der Entscheidung.“

Eingreifen und die Anfechtung sind die Vorstufe für die nachfolgende Geltendmachung der Rechte auf Darlegung des eigenen Standpunktes und auf Erwirkung der Überprüfung der Entscheidung. Aus Erwägungsgrund 71 der DSGVO sind diese Rechte jedoch „in jedem Fall" zu gewähren.

Mit der **Darlegung des eigenen Standpunktes** soll der betroffenen Person die Möglichkeit gegeben werden, die Besonderheiten des Einzelfalls aus ihrer Sicht vortragen zu können, die bei einer automatisierten Entscheidung keine Berücksichtigung finden würden. Martini und Nink erläutern dazu, dass die verantwortliche Stelle verpflichtet ist, die dargelegten Aspekte tatsächlich zu berücksichtigen, damit die Befugnis nicht zu einer „inhaltsleeren Floskel verkommt" (Martini & Nink 2017: 4). Die verantwortliche Stelle ist dann dazu angehalten, die Entscheidung zu überprüfen und sich mit den vorgebrachten Aspekten inhaltlich auseinanderzusetzen (ebd.). Allerdings sind die Rechte der Erwirkung eines direkten Eingreifens und der Darlegung des eigenen Standpunktes dahingehend geschwächt, dass Betroffene der Situation und möglichen Schädigungen durch automatisierte Entscheidungen gewahr sein müssen.

6.2.3.3 Informationspflichten

Aus dem Recht auf Darlegung des eigenen Standpunktes erwächst die Notwendigkeit, dass die betroffene Person des Entscheidungsverfahrens vorab oder während des Entscheidungsprozesses mit Informationen in derjenigen Detailliertheit versorgt werden muss, die ausreichend ist, damit die betroffene Person überhaupt sinnvoll dazu Stellung nehmen kann. Das wird durch die Informationspflichten des Art. 13 Abs. 2 lit. f und Art. 14 Abs. 2 lit. g DSGVO sichergestellt: „Danach muss der Verantwortliche frühzeitig sowohl über das Bestehen einer automatisierten Entscheidungsfindung informieren als auch aussagekräftige Informationen über die involvierte Logik sowie die Tragweite und angestrebte Auswirkungen einer solchen Verarbeitung für die betroffene Person bereit stellen [...]. Nach Art. 12 Abs. 1 müssen diese Informationen zudem in präziser, transparenter, verständlicher und leicht zugänglicher Form in einer klaren und einfachen Sprache erteilt werden [...]" (Scholz 2019: DSGVO Art. 22 Rn. 58).

Für die **„involvierte Logik"** präzisiert Scholz: „Unter dem Begriff „Logik" sind Angaben über Aufbau, Struktur und Ablauf der automatisierten Datenverarbeitung zu verstehen [...]. Die Information muss daher die tragenden Funktionsprinzipien der Anwendungsprogramme und die grundlegenden Entscheidungsmaßstäbe umfassen. Die technischen Einzelheiten

der verwendeten (Analyse-)Software oder der Quellcode müssen hingegen nicht mitgeteilt werden. Insoweit können sich die Verarbeiter in aller Regel auf den Schutz ihrer Geschäfts- und Betriebsgeheimnisse berufen. [...] Die betroffene Person muss aber verstehen können, in welcher Weise bestimmte Bewertungen und Klassifizierungen abgeleitet werden und welche Bedeutung und Gewichtung diese Werte für die automatisierte Entscheidung haben." (Scholz 2019: DSGVO Art. 22 Rn. 54).

Ähnlich äußert sich dazu Bäcker: „Diese Information bezieht sich auf die Methoden und Kriterien der Datenverarbeitung, etwa die Funktionsweise des Algorithmus, der bei der Bildung eines Scorewerts genutzt wird." (Bäcker 2018: Art. 13 Rn. 53–54). Auch Hoeren und Niehoff (2018: 56f.) argumentieren für die Offenlegung des Algorithmus in Form der Darlegung der „[...] Handlungsvorschriften und Programmabläufe mit den entsprechenden Gewichtungen [...]" (ebd., S. 57). Dies würde auch keine Gefahr für die Geschäftsgeheimnisse[115] der verantwortlichen Stelle bedeuten, da für eine solche Gefährdung auch der Quellcode verfügbar sein müsste. Zu den besonderen Formen der automatisierten Entscheidung mit Systemen der künstlichen Intelligenz führen sie aus, dass die tatsächlichen Entscheidungskriterien oft bei KI-Verfahren nicht nachvollziehbar sind. Es ist jedoch mit technischen Zusatzverfahren möglich, die besonders entscheidungsrelevanten Kriterien ausgeben zu lassen, die dann den Betroffenen dargelegt werden können. Allerdings handelt es sich dabei um Annährungswerte, sodass die Betroffenen auch über die Unwägbarkeiten der Verfahren aufgeklärt werden müssten (Hoeren & Niehoff 2018: 57–60). Weitergehende Informationen und Ansprüche, die eine detaillierte Erklärung im jeweiligen Einzelfall einer Entscheidung bedeuten würden, sind nach Wischmeyer (2018: 51f.) aus den Vorgaben der DSGVO nicht abzuleiten.[116]

Dennoch ist der Artikel für die Antidiskriminierung insofern wertvoll, da er klarstellt, dass neben der Pflicht über die Funktionsweise bzw. „Logik" zu informieren, auch über „die Tragweite und die angestrebten Auswirkungen einer derartigen Verarbeitung" (Art. 13 Abs. 2 lit. f DSGVO und Art. 15. Abs. 1

[115] Dabei beziehen sich die Autoren (ebd., S. 56f.) auf das sogenannte SCHUFA-Urteil des Bundesgerichtshofs (BGH), mit dem eine Klage auf Offenbarung der Scoreformel, die das Unternehmen SCHUFA für die Berechnung der Kreditwürdigkeit benutzt, mit der überwiegenden Bedeutung des Geschäftsgeheimnisses gegenüber Transparenzanforderungen abgewiesen wurde (BHG Urteil vom 28.1.2014, BGH (2014)).

[116] Zur Diskussion um ein sogenanntes „Recht auf Erklärung" siehe Goodman & Flaxman (2017), Wachter, Mittelstadt & Floridi (2017), Wischmeyer (2018), Edwards & Veale (2017).

lit. h DSGVO) informiert bzw. Auskunft gegeben werden muss. Danach muss eine verantwortliche Stelle beschreiben, „[...] worüber aufgrund der Datenverarbeitung entschieden werden soll, welche Entscheidungsmöglichkeiten bestehen und welche Verarbeitungsergebnisse zu welcher Entscheidung führen oder führen können." (Bäcker 2018: Art. 13 Rn. 55).

Schlussfolgerung

Angesichts einer derartigen Informationspflicht sollte beim Vorliegen von automatisierten Entscheidungen in Zukunft nicht nur über die Entscheidungsregeln, sondern auch über die Tragweite und Auswirkungen der Differenzierungsentscheidung, einschließlich Diskriminierungsrisiken, informiert werden müssen. Da die Information ex ante, also vor der Entscheidung erfolgen muss, hätten die potenziell von den Risiken Betroffenen die Möglichkeit, die Einwilligung zu verweigern. Des Weiteren hat dies den Effekt, dass sich die verantwortliche Stelle überhaupt erst einmal mit den Diskriminierungsrisiken auseinandersetzen muss, um darüber informieren zu können.

Ferner wäre zu prüfen, ob eine derartige Informationspflicht über die Tragweite und Auswirkungen der automatisierten Entscheidungen eine effektive Ergänzung zu unzureichenden Auskunftsansprüchen nach dem AGG, wie z. B. bei Bewerbungssituationen im Personalbereich, sein kann.

6.2.3.4 Kritik und Weiterentwicklungsbedarf

Insgesamt kritisiert Scholz die Regelungen des Art. 22 DSGVO: „Weder regelt die Vorschrift die äußerst relevante Frage, ob und unter welchen Voraussetzungen ein personenbezogenes Profil erstellt und verwendet werden darf, noch macht sie allgemein Vorgaben für einen nichtdiskriminierenden und transparenten Einsatz von Algorithmen." (Scholz 2019: DSGVO Art. 22 Rn. 8–11). Ebenso kritisch äußert sich Martini dahingehend, dass der Gesetzgeber ein größeres Gewicht auf die Ausschöpfung von Wertschöpfungspotenzialen und wirtschaftlichen Innovationen gelegt habe als auf den Schutz der Privatsphäre. Es werden nur äußerste Verbotsgrenzen für automatisierte Entscheidungen bestimmt, die „den Einzelnen zum bloßen Objekt einer ohne menschliches Eingreifen erfolgenden algorithmenbasierten Analyse machen" (Martini 2018: DS-GVO Art. 22 Rn. 8).

Des Weiteren sollte überdacht werden, dass, wenn algorithmische Verfahren von der verantwortlichen Stelle eingesetzt werden und nicht mehr nachvollzogen werden kann, wie die Entscheidung zustande kommt, dies quasi wie eine ausschließlich alleinig automatisierte Entscheidung zu interpretieren ist, selbst, wenn eine menschliche Person bei der Entscheidungsfindung involviert wäre. In diesen Fällen würde das Verbot greifen. Damit wäre zu überdenken, ob die in der DSGVO ausgeführten Kriterien, nach denen eine ausschließlich automatisierte Entscheidung vorliegt, auch um die Fähigkeit der Entscheidenden, die Entscheidungsempfehlungen des Computersystems nachzuvollziehen und gegenüber den Betroffenen erklären zu können, ergänzt und konkretisiert werden sollten.

Zusammenfassung und Schlussfolgerungen

- Insgesamt bleiben viele entscheidende Punkte des Verbots von automatisierten Entscheidungen nach der DSGVO rechtlich nicht eindeutig geklärt, insbesondere der Umfang und die Reichweite der Ausnahmen, vor allem bei der Verwendung geschützter Merkmale sowie die konkreten Informationspflichten. Zwar sieht die Rechtsnorm vor, dass Informationen in präziser, transparenter, verständlicher und leicht zugänglicher Form in einer klaren und einfachen Sprache erteilt werden müssen. Dies bezieht sich auch auf die sogenannte involvierte Logik, die als Aufbau, Struktur und Ablauf der automatisierten Datenverarbeitung interpretiert wird. Darüber hinaus müsste auch über die Tragweite und die angestrebten Auswirkungen einer derartigen Verarbeitung informiert werden, die auch über die Risiken, einschließlich Diskriminierungsrisiken, informieren müsste. Hierzu wären weitere verbindliche und vorab gelieferte Klärungen hilfreich, die nicht erst mit Gerichtsurteilen im Schadensfall die Auslegungsunsicherheiten mindern.

- Bei den Kriterien, wann das Verbot greift, sollte überdacht werden, ob mangelnde Nachvollziehbarkeit und Erklärbarkeit von „Entscheidungen“ der Computersysteme durch die verantwortliche Stelle zum Kriterium für das Vorliegen einer automatisierten Entscheidung konkretisiert wird.

- Als besonders problematisch kann gesehen werden, dass automatisierte Entscheidungen auch basierend auf geschützten

Merkmalen erlaubt sind, wenn Betroffene ihre ausdrückliche Einwilligung dazu gegeben haben. An der Funktionsfähigkeit der datenschutzrechtlichen Einwilligung wachsen allerdings zunehmend Zweifel.[117]

6.2.4 Kommunikative Prozesse bei Differenzierungsentscheidungen

Aus der Begründung des Schutzziels der freien Entfaltung der Persönlichkeit und dem Recht auf Selbstdarstellung[118] müssten Differenzierungsentscheidungen, wenn sie die Persönlichkeitsentfaltung betreffen, als kommunikative Prozesse gestaltet werden (Trute 1998: 825; Britz 2008: 185). Die Risiken der statistischen bzw. algorithmischen Diskriminierung könnten durch kommunikative Prozesse bei Differenzierungsentscheidungen gemildert werden. Demnach sollen statt einer nur einseitigen Beurteilung von Personen und der Zuschreibung von Merkmalen die kommunikativen Prozesse die Möglichkeit bieten, gemäß dem Recht auf Selbstdarstellung das eigene Selbstbild in den Prozess der Generierung des Bildes einer Person einfließen lassen zu können und Möglichkeiten des Abgleichs und der Korrektur der Fremdbilder zu schaffen.

Weder das Antidiskriminierungsrecht noch das derzeit im Datenschutzrecht operationalisierte Recht der informationellen Selbstbestimmung[119] schaffen jedoch eine ausreichende Grundlage dafür, dass in Entscheidungssituationen Betroffene immer eine Chance haben, dass das Persönlichkeitsbild in einem kommunikativen Prozess gebildet wird. Das sowohl im Diskriminierungs- als auch im Datenschutzrecht vorherrschende Vorgehen, erst nach der Wahrnehmung eines Schadensfalls oder bei entdeckten Fehlern Möglichkeiten zu haben, mit aufwendigen Selbstschutzmaßnahmen oder Rechtsprozessen gegebenenfalls lediglich eine Korrektur an Fremdbildern zu erreichen, kann dem Recht auf Selbstdarstellung für die Wahrung des Rechts auf freie Entfaltung der Persönlichkeit nicht entspre-

[117] Siehe dazu Abschnitt 6.1.3.1, S. 106.
[118] Siehe dazu Abschnitt 5.2.5, ab S. 93.
[119] Auch trotz des Rechts auf Einbringung des eigenen Standpunktes bei automatisierten Entscheidungen, siehe im Abschnitt 6.2.3.2, ab S. 118.

chen.[120] Veranschaulicht wird dies u.a. am Beispiel 22 (S. 50) zur Mehrfachdiskriminierung bei der Onlinekreditvergabe, bei der die betroffene Person keine Chance hatte, vor dem Kreditentscheid Hinweise auf ihre eigentlich vorhandene Zahlungsfähigkeit einzubringen.

Grundlage für kommunikative Prozesse ist die **Verständlichkeit und Nachvollziehbarkeit der Entscheidungsregeln**, die die Kriterien der Entscheidung und die Beziehungen zwischen den Kriterien sowie die Folgerungen daraus beinhalten. Dies ist notwendig, damit die betroffene Person überhaupt weiß, ob und auf welche Weise die Kriterien für ihre Lebenssituation zutreffend sind, und ob in ihrer Lebenssituation nicht ganz andere Umstände bzw. Kriterien das Differenzierungsziel erfüllen und sie somit erkennen kann, dass dies der entscheidenden Stelle kommuniziert werden muss. Die betroffene Person muss auch eine Chancen haben, nicht nur erkennen zu können, dass grundsätzlich die Möglichkeit der Darlegung des eigenen Standpunktes besteht, sondern wann und warum es wichtig ist, diesen einzubringen. In Entscheidungsverfahren müsste die Möglichkeit eröffnet werden, dass die betroffene Person auch für die Relativierung, Ergänzung, Anpassung oder Revision von Kriterien innerhalb von kommunikativen Entscheidungsverfahren sorgen kann, bevor eine Entscheidung getroffen wird. Diese grundsätzliche Möglichkeit, die Nachteile des durch Algorithmen geförderten Phänomens der statistischen Diskriminierung zu mindern, sollte erhalten bleiben und auch auf elektronischen Wegen umgesetzt werden.

Um die Effizienzgewinne zu wahren, können kommunikative Prozesse auch automatisiert gestaltet sein. Statt Automatisierung als die automatisierte Erzeugung von Fremdbildern über ein Individuum mit sogar quasi **heimlichem**[121] Entscheidungsvollzug über ein Individuum hinweg zu betreiben, kann auch nach IT-gestützten Erleichterungen der Darlegung des eigenen Standpunktes bzw. der Selbstdarstellung gesucht werden. Ebenso

[120] Erschwert wird dies noch dadurch, dass viele Angebote mit nicht verhandelbaren Vertrages- und Datenverarbeitungsbedingungen, die einseitig durch die Anbietenden gesetzt werden, erfolgen, zu denen die Betroffenen zustimmen oder auf den Dienst oder das Produkt ganz verzichten müssen.

[121] Grundsätzlich werden nach dem Recht der informationellen Selbstbestimmung der heimlichen Datenerhebungen und -verarbeitungen besondere Gefahren zugeordnet. „Die Heimlichkeit nimmt der betroffenen Person zum einen die Möglichkeit, sich der Informationspreisgabe und den damit verbundenen Nachteilsrisiken durch Verhaltensanpassung selbstschützend zu entziehen. Zum anderen ist die Möglichkeit nachträglichen Rechtsschutzes, insbesondere auch der nachträglichen Korrektur unzutreffender Informationen ausgeschlossen." Britz (2010: 579).

kann die Einbringung des Selbstbildes durch Möglichkeiten der Selbstselektion bzw. eigenen Zuordnung der betroffenen Personen zu Differenzierungskategorien verbessert werden. Statt der heimlichen Identifikation und Einordnung in Segmenten der Kundschaft, wären offen nachvollziehbare differenziertere Personentypisierungen oder Angebotskategorien von Produkten und Diensten, zu denen sich Kund*innen etc. zuordnen können, im Hinblick auf den Persönlichkeitsschutz sinnvoller.

6.2.5 Gestaltung von Onlineplattformen

Grundsätzlich kann angenommen werden, dass Onlineplattformen über algorithmische Systeme auch die dort auftretenden Diskriminierungsrisiken (siehe Abschnitt 5.1.3) einschränken könnten, und dies sogar besser als dies in konventionellen Märkten und Austauschbeziehungen der Fall sein kann. Die Betreibenden von Onlineplattformen können potenziell bei den auf der Plattform ablaufenden Interaktionen und Transaktionen unnötige oder unerwünschte Informationsflüsse, wie z. B. über bestimmte Merkmale von Personen (z. B. Geschlecht oder ethnische Herkunft), zentral und effizient steuern (Edelman & Luca 2014: 10) und als neutraler Intermediär handeln (Hannák u. a. 2017: 1914). Während beim konventionellen Handel und Austausch von Angesicht zu Angesicht die direkte Sichtbarkeit der Merkmale von Personen stereotypes Handeln befördern kann, könnten auf Onlinemärkten auch sensible bzw. besonders diskriminierungsanfällige Merkmale prinzipiell verborgen werden.[122]

Am Beispiel der Onlineplattform Airbnb machen Edelman u. a. **Verbesserungsvorschläge**, wie Onlineplattformen Diskriminierungsrisiken vermindern können, z. B. durch das Unterbinden der Anzeige des Namens von Teilnehmenden oder das Vermeiden von Überprüfungen der Personen vor der Buchung (Edelman u. a. 2017: 117ff.). Ähnlich schlagen Hannák u. a. für Onlinemarktplätze der Arbeitsvermittlung vor, dass die Marktplätze ohne demografische Angaben funktionieren sollen, also Nachfragen nach Arbeitsleistungen nur noch an irgendjemand und an selektierte Gruppen gerichtet werden sollten. Ferner meinen sie, dass voreingenommene Bewertungen mit Nachjustierungen durch die Betreibenden der Onlineplattformen aus-

[122] Levy und Barocas (2017) zeigen weitere Möglichkeiten und Beispiele der Gestaltung von Onlineplattformen auf, mit denen diskriminierendes Verhalten durch Nutzende vermindert werden kann.

geglichen werden könnten (Hannák u. a. 2017). Auch Chen u. a. schließen aus ihren Untersuchungsergebnissen (siehe Beispiel 3, S. 36), dass Onlinearbeitsmärkte eine aktive Rolle in der Überwindung von strukturellen Ungleichheiten auf Arbeitsmärkten spielen könnten, indem ihre Rangfolgealgorithmen Ergebnisse nach dem Kriterium der Gruppenfairness, d. h. entsprechend der Verteilung der relevanten Gruppierungen (z. B. Frauen und Männer) in der Bevölkerung präsentieren und nicht strukturelle Ungleichheiten widerspiegeln lassen (Chen u. a. 2018: 10). Auch an der Gestaltung der algorithmenbasierten Differenzierungsregeln von Onlineplattformen können Aufgaben der Antidiskriminierungsstellen ansetzen (s. u.).

6.3 Möglichkeiten der Antidiskriminierungsstellen

6.3.1 Auftrag und Kompetenzen

Europäische Antidiskriminierungsrichtlinien verpflichten die Mitgliedstaaten, Stellen zu benennen, deren Aufgabe es ist, die Verwirklichung des Grundsatzes der diskriminierungsfreien Gleichbehandlung aller Personen zu fördern[123] und darüber hinaus zu analysieren, zu beobachten und zu unterstützen.[124] Der deutsche Gesetzgeber hat sich dafür entschieden, mit der Antidiskriminierungsstelle des Bundes eine zentrale Stelle zur Umsetzung dieser Richtlinienvorgaben einzurichten. Die einschlägigen Richtlinien sehen drei Kernaufgaben vor, die Deutschland mit den §§ 25 ff. AGG, insbesondere § 27 AGG, umgesetzt hat: (1) von Diskriminierung Betroffene auf unabhängige Weise dabei zu unterstützen, ihrer Beschwerde wegen Diskriminierung nachzugehen, (2) unabhängige Untersuchungen zum Thema der Diskriminierung durchzuführen und (3) damit zusammenhängend unabhängige Untersuchungen zu veröffentlichen und Empfehlungen vorzulegen.[125] Anders als in

[123] Nach Art. 13 RL 2000/43/EG.

[124] Nach Art. 12 2004/113/EG und Art. 20 RL 2006/54/EG.

[125] Des Weiteren fügt Art. 20 RL 2006/54/EG als Aufgabenbeschreibung den Austausch verfügbarer Informationen auf geeigneter Ebene mit entsprechenden europäischen Einrichtungen hinzu. Ebenso haben nach Art. 12 RL 2000/43/EG, Art. 14 RL 2000/78/EG, Art. 11 RL 2004/113/EG und Art. 22 RL 2006/54/EG die Mitgliedstaaten den Dialog mit geeigneten Nichtregierungsorganisationen zu fördern, die an den vom Anwendungsbereich der Richtlinien erfassten Diskriminierungskategorien ein rechtmäßiges Interesse haben und sich an der Bekämpfung von Diskriminierung beteiligen.

einigen anderen EU-Ländern (z. B. Vereinigtes Königreich, Belgien, Rumänien) sieht das AGG weder ein eigenes Klagerecht für die deutsche Antidiskriminierungsstelle noch ein Verbandsklagerecht vor. Sie verfügt überdies über keine eigenen Ermittlungs- oder Offenlegungsbefugnisse.

6.3.2 Möglichkeiten von Untersuchungen und Nachweisen

Es wäre zu prüfen, ob die zuvor beschriebenen technischen Transparenz- und Testinstrumente[126] für die Arbeit von Antidiskriminierungsstellen geeignet, nützlich oder notwendig sind. Bei dieser Frage ist zu berücksichtigen, dass die direkte Überprüfung von Algorithmen bzw. Computersystemen ein hohes Maß von Sachwissen der Informatik erfordert. Soll die Prüfung auf algorithmische Diskriminierungsrisiken von außen erfolgen, kann einiges dafürsprechen, ein derartiges Wissen zu konzentrieren. Ebenso bedeutend dürften jedoch auch das Wissen von Sachverhalten, der Handelnden und Betroffenen und ihren Interessen, der Verhältnisse der relevanten Sektoren bzw. Branchen, der bereits getroffenen Abwägungen, Regelungen und Vorgaben zu Differenzierungen sein. Insbesondere spielen Kenntnisse von vergangenen Ungleichbehandlungen, diskriminierungsgefährdeten Situationen und diskriminierungsgefährdeter Personengruppen eine Rolle sowie Praktiken und (wirtschaftliche) Motivationen zu Differenzierungen und potenziellem Missbrauch. Diese Expertise ist erforderlich, um z. B. die Annahmen, die bei der Auswahl und Anwendung von Algorithmen, Modellen und Kriterien der Differenzierung zugrunde liegen, verstehen und prüfen oder die Legitimität des Vorgehens hinterfragen bzw. bestätigen zu können. Neben der Wissensausstattung sind Fragen des rechtlich ermächtigten Zugangs zu den notwendigen Daten, Algorithmen und Systemen eine Voraussetzung.

Unabhängig von der Frage nach der direkten Inspektion und dem Zugriff auf die Systeme stehen Antidiskriminierungsstellen oder Forschenden „klassische“ empirische Untersuchungen und Diskriminierungsanalysen auch für die Anwendungen von Algorithmen zur Verfügung, die durch spezialisierte Algorithmen-Audits ergänzt werden können.[127] Diese untersuchen vor allem die **Ergebnisse und Konsequenzen** für die Betroffenen im

[126] Siehe oben Abschnitt 6.1.1, ab S. 99.
[127] Siehe oben Abschnitt 6.1.2, ab S. 102.

Sinne von Ungleichbehandlungen oder erzeugten Ungleichheiten, die aus algorithmen- und datenbasierten Differenzierungsentscheidungen resultieren. Grundsätzlich kann mit der zunehmenden Computerisierung und Vernetzung (bzw. „Digitalisierung") von Kommunikation, Interaktionen bzw. Transaktionen im privatwirtschaftlichen und öffentlichen Bereich erwartet werden, dass auch die dabei anfallenden Datenmengen für statistische Untersuchungen von Ungleichheiten und Ungleichbehandlungen wächst. Hier ist es eher eine Frage der Zugangsregelungen zu den Daten, inwieweit sie für Untersuchungen der Antidiskriminierungsstellen verfügbar gemacht werden können. Dabei wäre an besondere oder erweiterte Pflichten zur Informationsbereitstellung der Anwendende gegenüber anerkannten Antidiskriminierungseinrichtungen zu denken. Für den öffentlichen Bereich könnten dazu auch die Möglichkeiten beachtet und gegebenenfalls erweitert werden, die sich aus dem Informationsfreiheitsgesetz ergeben (z. B. Fink 2018).

Bei vielen Beispielen in Kapitel 4 wurden Onlineplattformen, z. B. für Wohnungsvermietungen oder Stellenanzeigen, anhand ihrer Ergebnisse und Konsequenzen und dadurch auftretende Ungleichbehandlungen untersucht. Sie verdeutlichen, dass Mittel wie empirische Untersuchungen und Auditstudien geeignet scheinen, auch Onlineplattformen mit ihren algorithmisierten Transaktions- und Entscheidungsregeln auf diskriminierende Praktiken zu untersuchen. Es kann vermutet werden, dass durch die im Onlinebereich möglichen Techniken, wie automatisiert wiederholte Abfragen bzw. Nutzungen von Webcrawlern oder Nutzungen mit fingierten Personen bzw. Konten[128], sogar Erleichterungen bei der Datenbeschaffung im Vergleich zu einem Offline-Vorgehen möglich sind. Einige Beispieluntersuchungen[129] haben zudem gezeigt, dass neben den Nutzenden auch die Algorithmen der Onlineplattformen Diskriminierungsrisiken hervorgerufen haben, was „von außen", ohne direkten Zugriff auf den Programmcode aufgedeckt wurde. Beispiel 26 (S. 54) der Fallstudie des Fahrdienstvermittlers Uber zeigt jedoch, dass die Untersuchung von Onlineplattformen auch vor Zugangsproblemen zu den relevanten Daten stehen kann.

Des Weiteren sind einige Algorithmen und Systeme des maschinellen Lernens und der künstlichen Intelligenz als Onlinedienste nutzbar und können mit verschiedenen Datensätzen getestet werden, wie im Beispiel 46

[128] Siehe oben Abschnitt 6.1.2, ab S. 102.
[129] Siehe z. B. Beispiele 17 (S. 46), 19 (S. 47) oder 20 (S. 48).

(S. 74) die Gesichtserkennungsdienste von Microsoft, IBM und Face++ oder im Beispiel 47 (S. 75) das Gesichtserkennungssystem von Amazon. Die direkte Inspektion des Algorithmus und des Programmcodes schien dabei nicht notwendig gewesen zu sein, sondern Ungleichbehandlungen oder Diskriminierungen wurden durch die Analyse der Ergebnisse der Onlinedienste identifiziert.

Aus den Beispielfällen lassen sich jedoch (noch) keine verallgemeinerungsfähigen Aussagen schließen, da sie zu heterogen sind und nicht aus einer systematischen Erhebung stammen.[130] Denn andere Beispiele zeigen, dass Ungleichbehandlungen und Diskriminierungen erst in Rechtsstreitigkeiten ermittelt und (teilweise) belegt werden konnten, bei denen das Verfahren der Datenauswertung und die Entscheidungskriterien (teilweise) offengelegt werden mussten (z. B. Beispiele 22, S. 50 und 31, S. 60). Das Beispiel 30 (S. 59) des Systems des österreichischen Arbeitsmarktservice zeigte zudem, dass Diskriminierungsrisiken durch die Publikation der Dokumentation über Berechnungsformeln einer öffentlichen Diskussion zugänglich gemacht wurden.

6.3.3 Erfahrungen und Vorschläge anderer Antidiskriminierungsstellen

Beim Umgang mit algorithmen- und datenbasierten Diskriminierungsrisiken bestehen bei Antidiskriminierungsstellen in den EU-Ländern bisher nur wenig Erfahrungen. Auch liegen bislang nur wenige konkrete Fälle von Diskriminierungen vor. Zu den Möglichkeiten und Befugnissen von Untersuchungen wurden einige EU-Antidiskriminierungsstellen befragt, die bereits Vorerfahrungen mit dem Thema aufweisen konnten:[131]

[130] Die vielen Forschungsergebnisse zu Onlineplattformen und Onlinesystemen können auch dadurch zustande gekommen sein, dass, im Vergleich zum direkten Zugriff auf den Programmcode, die Onlinesysteme und Onlineplattformen noch relativ gut zu untersuchen sind, weil sich etwa Probleme des direkten Zugriffs auf Algorithmen oder Programmcode nicht stellten (z. B. wegen Schutz von Geschäftsgeheimnissen oder Urheberrechtsfragen) oder der Forschungsaufwand mit statistischen Auswertungen noch vertretbar gewesen sein könnte.

[131] Hierzu wurden einige europäische Antidiskriminierungsstellen per E-Mail befragt.

- Es sind Kenntnisse darüber notwendig, wie Algorithmen funktionieren und in welchen Lebensbereichen und Verhältnissen sie angewandt werden und geschützte Personengruppen betreffen. Dabei sind weniger technische Werkzeuge gefragt, als vielmehr Personal mit Kenntnissen über Daten, Datennutzung und Antidiskriminierungsrecht. Statt Expertise in Informatik oder Datenwissenschaft innerhalb der Einrichtung könnte die Einrichtung von strukturellen Partnerschaften mit Computer- oder Datenwissenschaftler*innen profitieren, um Diskriminierungen zu vermeiden, Diskriminierungsfälle aufzudecken oder Beweise zu den Fällen zu erhalten. Auditstudien und Diskriminierungs-Testings werden für geeignet gehalten, um auch Ungleichbehandlungen mit Involvierung von Computersystemen und Algorithmen zu erkennen.[132]

- Am Beispiel eines Falles zu diskriminierenden Versicherungstarifen in der Tschechische Republik wird die Wichtigkeit des Zugang zu den statistischen Daten und versicherungsmathematischen Methoden der Versicherungsunternehmen betont. Ohne den Zugang und die Verpflichtung auf Herausgabe von Daten und Informationen auf Anfrage der tschechischen Antidiskriminierungsstelle sei eine Bearbeitung des Falles und die Beurteilung, ob eine Diskriminierung vorliegt, nicht möglich.[133]

- Im Fall von diskriminierungsverdächtigen Stellenanzeigen im „sozialen“ Onlinenetzwerk Facebook (siehe Beispiel 4, S. 37) wird betont, dass für die Betroffenen die Kenntnisnahme von der Ungleichbehandlung nicht möglich war, da sie die Anzeigen nicht sehen konnten. Nur dadurch, dass andere Personen, die die Anzeigen wahrnehmen konnten, die Antidiskriminierungsstelle informiert haben, konnte die Stelle den Fall bearbeiten. Da in dem Fall die geschützten Merkmale „Alter“ und „Geschlecht“ verwendet wurden, konnten relativ leicht Beweise entdeckt, untersucht und erbracht werden. Für ein umfassendes Verständnis und um fähig zu sein, rechtlich gegen die Verwendung bestimmter Algorithmen vorgehen zu können, würden Computerfachkräfte notwendig sein, um Systeme

[132] Angaben durch Mitarbeiter*in der belgischen Antidiskriminierungseinrichtung Unia per E-Mail an den Autor, März 2019.

[133] Angaben durch Mitarbeiter*in der tschechischen Antidiskriminierungseinrichtung Office of the Public Defender of Rights, per E-Mail an den Autor, April 2019.

und Algorithmen genau zu analysieren. Die genaue Analyse wäre auch nur dann möglich, wenn die Algorithmen vollständig für Untersuchungen zugänglich gemacht würden, was aber aufgrund des Schutzes der Betriebs- und Geschäftsgeheimnisse für unwahrscheinlich gehalten wird. Die in der Antidiskriminierungsforschung und Analyse von Diskriminierungsfällen üblichen statistischen Untersuchungen werden auch für die Untersuchung von Ungleichheiten, die durch Algorithmen und Computersysteme verursacht werden, für geeignet gehalten.[134]

- Statistiken und insbesondere Gleichstellungsdaten („equality data“) werden als Kernelement bei der Analyse von vorgebrachten Fällen von Diskriminierungen mit Algorithmen und Computersystemen angesehen. Dazu können auch Reformen des Rechtsrahmens erfolgen, mit denen strukturelle Probleme gemildert werden sollen, wie mangelnde Daten oder fehlende Konsequenzen bei ausbleibenden Reaktionen auf Anfragen der Antidiskriminierungsstelle. Veränderungen im Rechtsrahmen könnten beispielsweise umfassen: Pflicht zur Erfassung von Gleichbehandlungsdaten durch bestimmte öffentliche Einrichtungen, Pflicht zur Berichterstattung über automatisierte Entscheidungssysteme an Antidiskriminierungsstellen, Pflicht zur Kooperation zwischen Anwendenden der Systeme und den für Gleichbehandlungsdaten zuständigen Stellen sowie rechtliche Konsequenzen, wenn Anwendende die Daten nicht auf Anfragen an die Gleichbehandlungsstellen liefern.[135]

6.3.4 Präventives Vorgehen und Kooperationsmöglichkeiten

An vielen Stellen basieren die Rechtsinstrumente des Diskriminierungs- und Datenschutzes darauf, dass ein Schaden oder Unrecht schon eingetreten ist. Wie dargelegt können algorithmische Diskriminierungen auch unbemerkt stattfinden oder bewusst verschleiert werden, unintendiert über Korrelationen zu geschützten Merkmalen ablaufen oder der Nachweis von

[134] Angaben durch Mitarbeiter*in der dänischen Antidiskriminierungseinrichtung Danish Institute for Human Rights, per E-Mail an den Autor, März 2019.

[135] Angaben durch Mitarbeiter*in der slowenischen Antidiskriminierungseinrichtung Advocate of the Principle of Equality, per E-Mail an den Autor, April 2019.

Diskriminierungen kann extrem schwer sein. Für Entwickelnde und Anwendende von Algorithmen und Computersystemen bietet zudem der Rechtsrahmen unzureichend Orientierung bei der Gestaltung von diskriminierungsfreien Algorithmen und Systemen, da zu viele Unsicherheiten und Auslegungsspielräume enthalten sind, die erst mit Gerichtsurteilen, wenn überhaupt, geklärt werden könnten. Unsicherheiten über die Legalität kann Fehlinvestitionen verursachen oder Innovationen werden nicht realisiert. Daher spricht einiges für ein präventives Vorgehen. Dazu könnten die Antidiskriminierungsstellen oder Aufsichtsinstitutionen vielfältige Aufgaben übernehmen, brauchen jedoch dazu die passende Ausstattung.

Antidiskriminierungsstellen, die vom Autor befragt wurden, ob ein präventives Vorgehen bei Diskriminierungsrisiken durch die Verwendung von Algorithmen und Computersystemen besser geeignet oder sogar notwendig ist (wie z. B. Untersuchungen, die durch die Einrichtung initiiert werden), machten folgende **Vorschläge** dazu (Anmerkungen des Autors in Klammern):

- Die frühe Sensibilisierung der Entwickelnden und verantwortlichen Handelnden bei der Schaffung von Algorithmen sei essenziell, um rechtlich-ethische Standpunkte mit technischen Standpunkten zusammenzubringen. Dies kann z. B. auch dadurch erreicht werden, dass die Diversität unter den Arbeitnehmenden bei den Entwickelnden erhöht wird, um ein Denken und Argumentieren nur durch eine Mehrheitsgruppe zu vermeiden. Mit fortschreitender Expertise in dem Gebiet der Algorithmen und auf Basis möglicher Auditstudien kann eine Strategie der Sensibilisierung und Bewusstseinsschaffung über die Diskriminierungsrisiken durch Algorithmen wichtig werden. Neben dem präventiven Vorgehen sei auch die Fortsetzung der Arbeit an qualitativ hochwertigen Gleichbehandlungsdaten („equality data“)[136] bedeutend, denn die Qualität der Algorithmen kann auch durch das Vorliegen von nicht-verzerrten und genauen Daten abhängen, insbesondere durch die Daten, die diskriminierungsgefährdete Gruppen abbilden.[137] (Ergänzend zu den Aussagen der Antidiskriminierungsstelle Unia, sei darauf hingewiesen, dass Gleichbehandlungsdaten auch helfen können, Diskriminierungsrisiken in Bezug auf

[136] Siehe dazu auch Baumann, Egenberger & Supik (2018).
[137] Angaben durch Mitarbeiter*in der belgischen Antidiskriminierungseinrichtung Unia, per E-Mail an den Autor, März 2019.

bestimme Gruppen besser zu identifizieren. Das kann auch den Nachweis von Diskriminierung im Kontext von Algorithmen unterstützen.)

- Angesichts der zunehmenden Verwendung von Werkzeugen der automatisierten Entscheidung durch Anbietende von Gütern und Diensten wird der präventive Ansatz für wichtig gehalten, insbesondere um Wahrnehmung und Kenntnis über die ethische Handhabung zu vermitteln. Dies soll vor allem durch die Informierung der Personen und Unternehmen, die den Einsatz solcher Techniken bei ihren Geschäften und Praktiken vorbereiten, geschehen. Andererseits sollen auch Beschwerdemechanismen für die Betroffenen von Diskriminierungen durch automatische Entscheidungssysteme zugänglich sein.[138] (Ergänzend zu den Aussagen der Antidiskriminierungsstelle Advocate of the Principle of Equality kann angemerkt werden, dass auch im Arbeitsleben innerbetriebliche Beschwerdeverfahren nach § 13 AGG die Möglichkeit für Beschäftigte eröffnen sollten, Diskriminierungsrisiken durch Algorithmen anzusprechen.)

So bietet beispielsweise die belgische Antidiskriminierungsstelle Unia eine Webseite an, auf der Unternehmen anhand eines bei ihnen auszufüllenden Fragebogens relativ schnell überprüfen können, ob Probleme hinsichtlich Diversität und entsprechender Entwicklungen dorthin vorliegen („Quick scan of diversity").[139] Unter Berücksichtigung der besonderen Herausforderungen hinsichtlich Methoden, technischen Fragen, Arbeitsbelastung und ethischer Fragen könnte ein vergleichbares Werkzeug beispielsweise auch für Entscheidungen über den Einsatz von Algorithmen angeboten werden.[140]

Auch eine Zusammenarbeit von Antidiskriminierungsstellen und Onlineplattformen im Sinne eines präventiven Vorgehens scheint lohnend (siehe auch Abschnitt 6.2.5). Dabei kann nach Ausgleichen von einerseits dem Ziel der Plattformen, möglichst viele Informationen über Nutzende bereitzustel-

[138] Angaben durch Mitarbeiter*in der slowenischen Antidiskriminierungseinrichtung Advocate of the Principle of Equality, per E-Mail an den Autor, April 2019.

[139] Siehe Webseite in französischer Sprache: https://www.ediv.be/site/fr/ediv-quick-scan-non-discrimination-et-egalite-des-chances oder in niederländischer Sprache: https://www.ediv.be/site/nl/ediv-quickscan-non-discriminatie-en-gelijke-kansen (zuletzt abgerufen am 28.8.2019).

[140] Angaben durch Mitarbeiter*in der belgischen Antidiskriminierungseinrichtung Unia per E-Mail an den Autor, März 2019.

len, und andererseits den Schutz von Träger*innen geschützter Merkmale sicherzustellen, gesucht werden. Letzteres wäre nicht nur durch Vermeidung der direkten Merkmalsnutzung, sondern auch durch Vermeidung von Rückschlüssen über Ersatzvariablen und Korrelationen zu erreichen. Durch die Massenwirkung der Onlineplattformen könnten so nicht-diskriminierende Praktiken gleich für vergleichsweise viele Menschen umgesetzt werden. Das Beispiel 9 (S. 41) zeigt ein kooperatives Vorgehen der National Fair Housing Alliance (NFHA) mit dem Unternehmen Facebook, bei dem durch die NFHA ein Trainingsprogramm für das Unternehmen angeboten wird.

Des Weiteren schlägt Zuiderveen Borgesius die Kooperation zwischen Datenschutzeinrichtungen und Antidiskriminierungsstellen sowie für öffentliche Einrichtungen, die Algorithmen der künstlichen Intelligenz verwenden wollen, eine Pflicht zur vorherigen Konsultation von Antidiskriminierungsstellen und deren Beteiligung bei öffentlichen Beschaffungsprozessen vor (Zuiderveen Borgesius 2018: 31).

Schlussfolgerung

Ein auf Prävention ausgerichtetes, kooperatives Vorgehen zwischen Antidiskriminierungsstellen und Entwickelnden bzw. Anwendenden von Algorithmen und Computersystemen könnte mehrere Elemente umfassen, wie beispielsweise (1) Beratung über die legitime oder verbotene Verwendung von geschützten Merkmalen abhängig von Entscheidungssituationen und betroffenen Personengruppierungen, (2) Interpretation und Beratung zur Verwendung von Proxies, Ersatzinformationen bzw. -variablen mit Korrelationen zu geschützten Merkmalen bzw. von scheinbar neutralen Kriterien bei der mittelbaren Diskriminierung oder (3) Interpretation und Eruierung von Umsetzungsmöglichkeiten von Gerechtigkeits- und Fairnesskriterien für verschiedene Differenzierungssituationen.

6.3.5 Vorschläge für die Antidiskriminierungsstelle des Bundes

Aus den vorhergehenden Erkenntnissen sind die nachfolgenden Vorschläge für die Antidiskriminierungsstelle des Bundes (ADS) ableitbar:

- Antidiskriminierungsrechtlich relevant sind all jene algorithmen- und datenbasierten Differenzierungsentscheidungen, die zu Schlechterbehandlungen aufgrund von einem im AGG geschützten Merkmal in den Bereichen Arbeitsleben und Zugang zu Gütern und Dienstleistungen führen können. Das AGG verbietet bereits jetzt diese Diskriminierungen, unabhängig davon, ob sie mit Verwendung von Algorithmen getroffen werden oder nicht.[141] Ein weitergehendes gesetzliches Verbot erscheint in den Bereichen Arbeits- und Zivilrecht daher nicht notwendig. Allerdings sind die Entwicklungen bei Algorithmen, insbesondere der künstlichen Intelligenz, weiter zu beobachten und zu erforschen, um gegebenenfalls Anpassungen bei den geschützten Merkmalen des AGG vorzunehmen.

- Das AGG hat aber Schwächen, da es auf individuelle Rechtsdurchsetzung beschränkt ist. So kann eine diskriminierende Praxis nicht untersagt, sondern nur von einzelnen Betroffenen in zivilrechtlichen Rechtsstreiten Schadenersatz oder eine Entschädigung eingeklagt werden. Der Ansatz des lediglich punktuellen Vorgehens im Einzelfall erscheint mit Blick auf die gegebenenfalls systematische Schlechterbehandlung von vielen Betroffenen durch algorithmenbasierte Differenzierungen nicht sachgerecht. Das Verbandsklagerecht wäre eine notwendige Antwort, um dem Massenphänomen[142] von möglichen algorithmenbasierten Diskriminierungen und der schlechteren Wahrnehm- und Nachweisbarkeit von algorithmenbasierten Diskriminierungen gerecht zu werden.

- Um Diskriminierungen durch algorithmengestützte Entscheidungen (vor Gericht) nachweisen und Schadenersatz- bzw. Entschädigungsansprüche begründen zu können, sollten Dokumentationspflichten für Systeme mit künstlicher Intelligenz bzw. besonders diskriminierungslastige Systeme vorgeschrieben werden. In konkreten Verdachtsfällen sollte der Zugang der Antidiskriminierungsstelle zu derartigen Dokumentationen zur Überprüfung ermöglicht werden. Zugang oder Herausgabe der Dokumentation müssten für solche Fälle gesetzlich geregelt werden. Dabei wäre

[141] Siehe auch Abschnitt 6.2.1, S. 110.
[142] Siehe zum Massenphänomen S. 23.

zudem zu prüfen, ob sich Zugang oder Herausgabe auch auf Datensätze[143] und Algorithmen selbst beziehen müssten und wie dann der Geheimnisschutz gesichert werden kann.

- Aus Sicht der Betroffenen können Unklarheiten über Zuständigkeiten bei Beschwerden und Rechtsunterstützung entstehen, wenn für algorithmen- und nicht-algorithmenbasierte Differenzierungsentscheidungen unterschiedliche Zuständigkeiten und Verfahrenswege eingerichtet würden. Zumal diese Unterscheidung in der Praxis immer weniger präzise gemacht werden könnte. Unterschiedliche Zuständigkeiten wären auch zu vermeiden, da ansonsten unterschiedliche Schutzniveaus entstehen könnten.

- Aus dem gesetzlichen Auftrag, als nationale Gleichbehandlungsstelle für die Umsetzung der europäischen Antidiskriminierungsrichtlinien verantwortlich zu sein, resultiert eine Zuständigkeit für diskriminierungsrechtliche Untersuchungen und Bewertungen. Um den gesetzlichen Auftrag zu erfüllen, sollte technischer Sachverstand aufgebaut werden oder dieser durch die Kooperation mit Forschungseinrichtungen erlangt werden. Des Weiteren sollte dieser gesetzliche Auftrag bei der weiteren Gestaltung des Regulierungsrahmens von algorithmenbasierten Differenzierungen und Entscheidungssystemen berücksichtigt und eingebettet sein.

- Zur Erfüllung des gesetzlichen Forschungsauftrags sollte die Antidiskriminierungsstelle auch verdachtsunabhängige Überprüfungen (Testings oder Algorithmen-Audits) der Ergebnisse von Differenzierungsentscheidungen, wie z. B. bei Onlineplattformen, vornehmen. Auch hierzu könnte sie mit Forschungseinrichtungen zusammenarbeiten (s. o.).

[143] Die zahlreichen beschriebenen Beispielfälle, bei denen sich Diskriminierungsrisiken aus der (Weiter-) Verwendung von Datensätzen, die vorherige Ungleichbehandlungen abbilden, insbesondere bei der Bildung von Systemen der Risikobewertung (siehe etwa dazu Beispiele 27 oder 34), würden dafürsprechen. Ein grundsätzliches Verständnis der Praktiken, durch wen und wie die Daten erzeugt und ausgewertet werden, scheint notwendig, um auch die Diskriminierungsrisiken durch die Verwendung der Systeme abschätzen zu können. Allerdings setzt dies die in Abschnitt 6.1.2 beschriebenen hohen Kompetenzerfordernisse voraus.

— Die Antidiskriminierungsstelle sollte auch präventive Angebote zur Vermeidung von algorithmenbasierten Diskriminierungsrisiken machen (s. o.) und dafür mit Unternehmen zusammenzuarbeiten. Ferner scheinen verbindliche Vorgaben zur Konsultation der Antidiskriminierungsstelle bei der Beschaffung und vor dem Einsatz von diskriminierungsanfälligen Computersystemen, wie z. B. bestimmte KI-Systeme, durch öffentliche Stellen sinnvoll.

6.4 Bedarf nach gesellschaftlichen Abwägungen und Festlegungen

Über konkretere Handlungsbedarfe und -optionen im Sinne einer Optimierung bestehender Regulierungsstrukturen hinaus ergibt sich ein Bedarf nach gesellschaftlichen Abwägungen und politischen und rechtssetzenden Festlegungen, die grundsätzlich die Eignung bestehender regulatorischer und institutioneller Ansätze angesichts der Entwicklungen von Algorithmen, Anwendungen und Praktiken zur Diskussion stellen. Das betrifft den Ansatz des Selbstschutzes und der daraus resultierenden Lasten des Individuums und der Legitimität algorithmen- und datenbasierter Differenzierungen in Hinblick auf die Abwägung von Vorteilen und gesellschaftlichen Risiken.

6.4.1 Lasten der betroffenen Individuen

Sowohl das Recht auf informationelle Selbstbestimmung als auch der Antidiskriminierung legen Verantwortungslasten auf das betroffene Individuum, die unrechtmäßigen Datenverarbeitungen und ungerechtfertigten Ungleichbehandlungen festzustellen und dagegen vorzugehen. Es stellen sich jedoch Fragen, ob diese rechtlichen Grundkonzeptionen überhaupt noch geeignet sind, angesichts einer zunehmenden Menge an daten- und algorithmenbasierten sowie automatisierten Entscheidungsverfahren sowie deren besonderen Eigenschaften. Denn derartige Verantwortungslasten erfordern bei den betroffenen Individuen sehr hohe fachliche, kognitive und zeitliche Voraussetzungen, um (a) die vielen Situationen mit Datenverarbeitungen und Differenzierungen überhaupt wahrzunehmen, (b) gegebenenfalls die aus den datenschutzrechtlichen Informationspflichten resultierenden Informationen zu verarbeiten (wie z. B. die „involvierte

Logik“ bei automatisierten Entscheidungen) sowie Auskunfts-, Korrektur- oder Löschrechte durchzusetzen und vor allem um (c) die individuellen Konsequenzen, die aus Datenverarbeitungen und vielfältigen (potenziellen) Differenzierungsentscheidungen resultieren, für sich abzuschätzen und das Risiko möglicher Diskriminierungen für sich zu erkennen.

So führen z. B. viele Datenverarbeitungen erst nach langer Zeit zu Differenzierungsentscheidungen, was vom Individuum hohe prognostische Fähigkeiten erfordert. Die Konsequenzen aus Datenverwendungen für fremde Entscheidungszwecke und in anderen Kontexten sind ohnehin kaum vom Individuum abschätzbar. Die Skepsis wird zudem durch die zunehmend als unzulänglich wahrgenommene informierte Einwilligung[144], die die genannten Abschätzungen des Individuums im Zeitpunkt der Einwilligung erfordert, und durch das erschwerte Erkennen und Nachweisen von Diskriminierungen durch die (potenziell) Betroffenen[145] gestützt. Da eine Reihe von Erfassungen von personenbezogenen Daten, wie das Webtracking, beim Gebrauch von Smartphones und deren Apps oder die Auswertung der Kommunikation in „sozialen“ Onlinenetzwerken, mehr oder weniger heimlich vor den Nutzenden ablaufen, sind zudem die Möglichkeiten des Selbstschutzes stark einschränkt.

Die eingeschränkten Möglichkeiten des Selbstschutzes der (potenziell) betroffenen Individuen sind auch bei Vorschlägen zur Regulierung von algorithmischen Diskriminierungsrisiken zu beachten, wie etwa beim Vorschlag für eine Kennzeichnungspflicht bei der Verwendung von automatisierten Entscheidungen. Da dabei nur auf die Existenz von automatisierten Entscheidungssystemen, aber nicht auf deren Konsequenzen hingewiesen würde, würde das Instrument keine Verbesserung des Problems der unzureichenden Selbstschutzmöglichkeiten liefern.

Nach dem Subsidiaritätsprinzip können daraus Forderungen nach einem stärker repräsentativen Vorgehen von Schutzeinrichtungen wie Antidiskriminierungsstellen, Verbraucherschutz- und Datenschutzeinrichtungen oder spezialisierten Behörden folgen, entweder mit (weiteren) Hilfen für Individuen oder, was noch wirksamer scheint, einem Vorgehen anstelle von Individuen. Auch eine weitere Verantwortungsverlagerung für Vermeidungsmaßnahmen auf die Entwickelnden und Anwendenden von al-

[144] Siehe Abschnitt 6.1.3.1, ab S. 106.
[145] Siehe Abschnitt 6.1.3.2, ab S. 107.

gorithmen- und datenbasierten Differenzierungen (die aber der Aufsicht bedarf) wäre zu diskutieren. Dazu könnten beispielsweise die datenschutzrechtlichen Pflichten zur Dokumentation, zur Durchführung von Datenschutzfolgenabschätzungen, zur Berufung von Datenschutzbeauftragt*innen um Antidiskriminierungsvorgaben ausgebaut oder ergänzt werden.

Das Vorgehen durch Stellvertretende kann vor allem auch durch zunehmende strukturelle Überlegenheit[146] der Anwendenden und Anbietenden gegenüber den Betroffenen notwendig werden. Diese kann insbesondere durch Verstärkung von Angewiesenheiten auf ein bestimmten Dienst oder ein bestimmtes Produkt auftreten, die wiederum durch Netzwerkeffekte der Onlineplattformen oder sonstigen Monopolisierungstendenzen entstehen können oder dadurch erhöht werden, dass die Persönlichkeitseigenschaften eingehender ausgeforscht und ausgenutzt werden können. Strukturelle Überlegenheit kann auch nicht zuletzt dadurch erhöht werden, dass Ausweichmöglichkeiten in Form von offline oder analogen Alternativen reduziert werden.

Hinzu kommt das immer wieder vorgebrachte Problem, dass das Datenschutzrecht konzeptionell von der Einzelperson ausgeht, aber die algorithmenbasierten Differenzierungen sich häufig auf Gruppierungen beziehen, ohne dass unbedingt ein Individuum mit dem Namen oder auf andere Weise eindeutig identifiziert sein müsste. In Zweifelsfällen bzw. Rechtsstreitigkeiten um den Personenbezug könnte sich das Recht dadurch als „zahnlos“ erweisen (Barocas & Nissenbaum 2014; Mantelero 2016; Zuiderveen Borgesius 2016). Hierzu wären Klärungen durch eindeutige Rechtsvorgaben notwendig. Ebenso begründet die Problematik ein Vorgehen durch stellvertretende Institutionen anstelle der betroffenen Individuen.

6.4.2 Legitimität von Differenzierungen

Wie in Kapitel 5 beschrieben, erzeugen algorithmen- und datenbasierte Differenzierungen neben technischen Risiken auch gesellschaftliche Risiken. Bei gesellschaftlichen Risiken helfen technische Verbesserungen der Algorithmen oder Datensätze nicht, sondern die Vermeidung der gesellschaftlichen Risiken sind Aufgaben von gesellschaftlichen Abwägungen

[146] Siehe Abschnitt 5.2.6, ab S. 95.

und Festlegungen, die den Stellenwert gesellschaftlicher Werte und Ziele sowie die gewünschten oder unerwünschten Praktiken der Differenzierung bestimmen. Auch bisher war der Umgang mit statistischer Diskriminierung Gegenstand von gesellschaftlichen Abwägungen und Festlegungen, vor allem durch die Ausgestaltung des Antidiskriminierungsrechts. Doch eine Reihe von Entwicklungen bei Datenverarbeitungen und Algorithmenanwendungen erfordert eine Neubewertung von Vor- und Nachteilen durch gesellschaftliche Abwägungsprozesse:

(1) Die Legitimität der algorithmenbasierten Differenzierungen wird teilweise begründet mit Kostenabwägungen und Effizienzzielen, die den Gebrauch von Ersatzinformationen bzw. Proxies bei Differenzierungsentscheidungen vor allem mit der effizienten Lösung von Informationsdefiziten rechtfertigen. Die alternative Einzelfallprüfung kann bei vielen Differenzierungsentscheidungen zu kostspielig sein oder selbst zu Problemen des Schutzes der Privatsphäre oder zu Stigmatisierungen führen (Britz 2008). Ein Verzicht auf Formen der statistischen bzw. algorithmischen Differenzierung kann gesellschaftliche Kosten verursachen und zwar in Form von relativen Ineffizienzen bzw. des entgangenen instrumentellen Nutzens. Konkret würde sich das in Form der Kosten von Einzelfallprüfungen ausdrücken (Schauer 2018: 50).

Demgegenüber stehen die Risiken des Generalisierungsunrechts durch unangemessene oder fehlerhafte Ersatzinformationen.[147] So sind in Abwägungen die Erforderlichkeit und Angemessenheit der algorithmenbasierten Differenzierung zu berücksichtigen, genauso wie die Frage, ob nicht weniger risikoreiche Alternativen zur Verfügung stehen. Auch Aspekte, wie eindeutige Nachweise der durchgehenden Verbesserung der Genauigkeit der Prognosen oder der Objektivität der Entscheidungen, Festlegungen der für verschiedene Risiken, Verwendungskontexte und Lebensbereiche akzeptablen Fairness- und Fehlerraten[148] sowie Festlegungen zu wissenschaftlich anerkannten Verfahren[149] der Datenanalyse und algorithmenbasierten Entscheidungsverfahren sind in diesem Zusammenhang zu klären.

Bei den Abwägungen stellen sich jedoch vor allem auch Fragen, wie Effizienzgewinne und gesellschaftlichen Kosten bzw. Risiken der algorithmen-

147 Siehe Abschnitt 5.2.1, ab S. 86.
148 Siehe Abschnitt 6.1.1, ab S. 99.
149 Siehe Abschnitt 6.2.2, ab S. 113.

basierten Differenzierung in der Gesellschaft verteilt werden, z. B. ob sie allen zufallen oder nur wenigen und ob diejenigen, denen die Effizienzgewinne zufallen auch diejenigen sind, die die Risiken tragen oder aber die Risiken externalisiert werden. So ist beispielsweise zu bedenken, dass bei der statistischen Differenzierung durch Private eine Ressourcenschonung nicht gemeinnützig stattfindet, sondern lediglich im eigennützigen Interesse ist (Britz 2008: 49f.). Ferner ist die Verteilung der Effizienzgewinne auch mit Blick auf die zunehmende Konzentration der relevanten Märkte und die Dominanz weniger Unternehmen zu beurteilen.

(2) Dadurch, dass Daten über die Zugehörigkeit zu bestimmten Personenkategorien und Algorithmen sowie sie beinhaltende Softwaresysteme nun vergleichsweise leicht und kostengünstig zur Verfügung stehen, ist es wahrscheinlich, dass derartige Daten und darauf basierende Entscheidungsverfahren „übernutzt" werden (Schauer 2003). Daraus folgt, dass algorithmen- und datenbasierte Differenzierungen in vielen Bereiche realisiert werden, in denen zuvor Gleichbehandlung vorherrschte, oder es können sogar Einzelfallprüfungen und andere Entscheidungsregelungen systematisch zurückgedrängt werden, auch wenn sie mit vertretbarem Aufwand möglich wären:

(a) Algorithmenbasierte Differenzierungen können in Bereiche vordringen, bei denen Differenzierungen aus gesellschaftlichen Gleichheitszielen oder sozialpolitischen Zielen zuvor ungewollt waren.[150]

(b) Algorithmenbasierte Differenzierungen werden zunehmend für Entscheidungen eingesetzt, die die Menschenwürde und die freie Entfaltung der Persönlichkeit maßgeblich bestimmen (wie z. B. Freiheitsentzug, Ausmaß von staatlichen Kontrollen sowie Zugang zu Wohnungen, Arbeitsstellen, Ausbildungsmöglichkeiten oder Krediten). In den tatsächlichen Automatisierungspraktiken können Menschen zunehmend nur noch als bloßes Mittel behandelt werden, weil betroffene Personen effektiv nicht mehr die Möglichkeit haben, den Praktiken zuzustimmen.[151]

(3) Algorithmen- und datenbasierte Differenzierungen beruhen in vielen Fällen auf der Erfassung und Auswertung von großen Mengen (zunehmend

[150] Siehe Abschnitt 5.2.3, ab S. 90.
[151] Siehe Abschnitt 5.2.4, ab S. 91.

sensibleren) personenbezogener Daten, die das Recht der **informationellen Selbstbestimmung** unterminieren können. Die Schutzziele der freien Entfaltung der Persönlichkeit werden vor allem durch das Antidiskriminierungsrecht und das Recht auf informationelle Selbstbestimmung realisiert. Sie sind (nach Britz 2010)[152]:

(a) Gewährleistung der äußeren Entfaltungsfreiheit durch Sicherung der Verhaltensfreiheit und Schutz vor nachteiligen Entscheidungen anderer, indem die potenziell verhaltenseinschränkenden Entscheidungen anderer durch die Betroffenen beeinflussbar bleiben, um für sie möglichst günstig auszufallen,

(b) Gewährleistung der inneren Entfaltungsfreiheit dadurch, dass die Persönlichkeitsentwicklung in interaktiven Prozessen noch als eigene und freie durch das Individuum wahrgenommen werden kann und das Individuum sich gegen Fremdbilder noch behaupten kann sowie

(c) Schutz der Unbefangenheit des individuellen Verhaltens durch Minderung der freiheitshemmenden Effekte einer „abstrakten Ungewissheit" bzw. Vermeidung von Einschüchterungseffekten. Letztere entstehen durch Informationsbestände und Verarbeitungszwecke, die für das einzelne Individuum nicht mehr überschaubar sind.[153]

So können beispielsweise algorithmenbasierte Differenzierungen zunehmend auf umfassende und detaillierte Persönlichkeitsprofile[154] beruhen, die als gesteigerte Form der Fremdbestimmung geeignet sind, einer Person ein Fremdbild „überzustülpen", ohne dass die betroffene Person eine Chance hat, die eigene Persönlichkeit und Rolleninterpretation in sozialen Kon-

[152] Siehe auch Abschnitt 5.2.5, ab S. 93.

[153] Diese Einschüchterungseffekte werden nicht nur durch unüberschaubare Datenbestände sondern auch durch die Zweckentfremdung von personenbezogenen Daten erwartet. So können Menschen von der Nutzung eigentlich vorteilhafter Anwendungen absehen, wenn sie befürchten müssen, dass die bei der Nutzung der Anwendung erhobenen, auf sie bezogenen Daten in anderen Kontexten und für andere Zwecke verwendet werden können, die nicht in ihrem Interesse sind; vgl. Yeung (2018: 33). Diese Einschüchterungs- und Selbsteinschränkungseffekte werden auch unter dem Begriff „chilling effects" untersucht; siehe z. B. Baruh (2007), Schwartz (1999), Das & Kramer (2013), Lang & Barton (2015), Marder u. a. (2016), Marthews & Tucker (2017), Penney (2016), (2017), Orwat & Schankin (2018).

[154] Siehe Abschnitt 2.2.2, ab S. 7.

texten zu entwickeln.[155] Daher sind auch bedenkliche Entwicklungen im Datenschutz näher in ihren Auswirkungen auf den Schutz vor algorithmen- und datenbasierten Diskriminierungen zu prüfen: Dazu zählen die problematische Zusammenführung von personenbezogenen Datensätzen, insbesondere durch den Datenhandel bzw. das Daten-Brokerage oder die (unternehmensinterne) Datenzusammenführung sowie die Aufweichung der Zweckbindung der Datenverwendung[156] und die Übertragung auf neue Verwendungszwecke. Des Weiteren sind bei gesellschaftlichen Abwägungen zu berücksichtigen, wem (den Anwendenden oder Betroffenen) die Vorteile der gesteigerten Überwachung, Datenauswertungen und Differenzierungen zufallen und auf wen Nachteile externalisiert werden.

Abwägungen, Festlegungen und Umsetzungen in regulatorischen Maßnahmen sollten sich für einzelne Entscheidungssituationen und Kontexte jeweils nach dem Ausmaß des Risikos des Generalisierungsunrechts, nach dem Ausmaß der Gefährdung der Menschenwürde und der freien Persönlichkeitsentfaltung richten, ebenso nach den Möglichkeiten und Grenzen des Selbstschutzes und der möglichen Überlastung der Individuen. Instrumentell kann dies die detaillierte Regulierung von Entscheidungsregeln[157] umfassen und auch – je nach Risikoausmaß – Verbote[158] oder eine Regulierungen der Verwendung bestimmter Algorithmen, Datenverarbeitungs- und Entscheidungsverfahren, Differenzierungsformen oder Entscheidungskriterien enthalten. Weitere Instrumente können die rechtliche Erforderlichkeit zur Überprüfung von gesellschaftlichen Risiken für Gleichheitsziele und die Persönlichkeitsentfaltung durch Entwickelnde und Anwendende oder deren Verpflichtung zu Schutzmaßnahmen im Sinne der Diskriminierungsvermeidung sein und bis hin zur Stärkung der Kompetenzen und Befugnisse von spezialisierten Einrichtungen des Diskriminierungs- und Datenschutzes reichen.

---

155 Siehe dazu Abschnitt 5.2.5, ab S. 93.

156 Siehe zur Zweckbindung in der DSGVO Raabe & Wagner (2016).

157 Siehe Abschnitt 6.2, ab S. 110.

158 Ein Bespiel für ein Verbot von algorithmischen Systemen ist das Verbot von Gesichtserkennungssystemen, das z. B. durch die Stadt San Francisco verhängt wurde. Vgl. Conger, Fausset & Kovaleski (2019).

7. Literaturverzeichnis

Acquisti, Alessandro; Taylor, Curtis; Wagman, Liad (2016): The Economics of Privacy, in: Journal of Economic Literature, 54. Jg., H. 2, S. 442–492.

Agrawal, Ajay; Gans, Joshua; Goldfarb, Avi (2016): The Simple Economics of Machine Intelligence, in: Harvard Business Review, 17. Jg., H. Nov., S. 2–5.

Agrawal, Ajay; Gans, Joshua; Goldfarb, Avi (2018): Prediction Machines. The Simple Economics of Artificial Intelligence; Boston: Harvard Business Review Press.

AI Now Institute (2018): Litigating Algorithms: Challenging Government Use of Algorithmic Decision Systems. An AI Now Institute Report, in collaboration with Center on Race, Inequality, and the Law and Electronic Frontier Foundation; New York: New York University, AI Now Institute.

Albers, Marion (2017): Informationelle Selbstbestimmung als vielschichtiges Bündel von Rechtsvorschriften und Rechtspositionen, in: Michael Friedewald, Jörn Lamla und Alexander Roßnagel (Hrsg.): Informationelle Selbstbestimmung im digitalen Wandel; Wiesbaden: Springer, S. 11–35.

Ali, Muhammad; Sapiezynski, Piotr; Bogen, Miranda; Korolova, Aleksandra; Mislove, Alan; Rieke, Aaron (2019): Discrimination through optimization: How Facebook's ad delivery can lead to skewed outcomes, arXiv e-prints.

Allhutter, Doris (2019): AMS-Algorithmus am Prüfstand. ITA-Dossier Nr. 43; Wien: Institut für Technikfolgen-Abschätzung (ITA).

Alpaydin, Ethem (2016): Machine Learning; Cambridge, London: The MIT Press.

Altenburger, Kristen M.; Ho, Daniel E. (2018): When Algorithms Import Private Bias into Public Enforcement: The Promise and Limitations of Statistical Debiasing Solutions, in: Journal of Institutional and Theoretical Economics (JITE), 175. Jg., H. 1, S. 98–122.

an der Heiden, Iris; Wersig, Maria (2017): Preisdifferenzierung nach Geschlecht in Deutschland – Forschungsbericht. Eine Studie im Auftrag der Antidiskriminierungsstelle des Bundes; Baden-Baden: Nomos.

Ananny, Mike; Crawford, Kate (2018): Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability, in: New Media & Society, 20. Jg., H. 3, S. 973–989.

Angrave, David; Charlwood, Andy; Kirkpatrick, Ian; Lawrence, Mark; Stuart, Mark (2016): HR and analytics: why HR is set to fail the big data challenge, in: Human Resource Management Journal, 26. Jg., H. 1, S. 1–11.

Angwin, Julia; Larson, Jeff; Mattu, Surya; Kirchner, Lauren (2016): Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks, in: ProPublica, Onlineartikel vom 23.5.2016, abrufbar unter: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (Letzter Zugriff am 27.8.2019).

Angwin, Julia; Scheiber, Noam; Tobin, Ariana (2017): Dozens of Companies Are Using Facebook to Exclude Older Workers From Job Ads, in: ProPublica, Onlineartikel vom 20.12.2017, abrufbar unter: https://www.propublica.org/article/facebook-ads-age-discrimination-targeting (letzter Zugriff am 28.8.2019).

Angwin, Julia; Tobin, Ariana; Varner, Madeleine (2017): Facebook (Still) Letting Housing Advertisers Exclude Users by Race; in: ProPublica, Onlineartikel vom 17.11.2017, abrufbar unter: https://www.propublica.org/article/facebook-advertising-discrimination-housing-race-sex-national-origin (Letzter Zugriff am: 28.8.2019).

Arentz, Christine; Rehm, Rebekka (2016): Behavior-based Tariffs in Health Insurance - Compatibility with the German System; Cologne: Otto Wolff Institut für Wirtschaftsordnung.

Arrow, Kenneth J. (1973): The Theory of Discrimination, in: Orley Ashenfelter und Albert Rees (Hrsg.): Discrimination in Labor Markets; Princeton: Princeton University Press, S. 3–33.

Arrow, Kenneth J. (1998): What has economics to say about racial discrimination?, in: Journal of Economic Perspectives, 12. Jg., H. 2, S. 91–100.

Bäcker, Matthias (2018): DS-GVO Art. 13 Informationspflicht bei Erhebung von personenbezogenen Daten bei der betroffenen Person, in: Jürgen Kühling und Benedikt Buchner (Hrsg.): Datenschutz-Grundverordnung, Bundesdatenschutzgesetz: DS-GVO/BDSG, Kommentar, 2. Aufl.; München: Beck.

Barocas, Solon; Hood, Sophie; Ziewitz, Malte (2013): Governing Algorithms: A Provocation Piece; New York: New York University, Department of Media, Culture, and Communication.

Barocas, Solon; Nissenbaum, Helen (2014): Big Data's End Run around Anonymity and Consent, in: Julia Lane, Victoria Stodden, Stefan Bender und Helen Nissenbaum (Hrsg.): Privacy, Big Data, and the Public Good. Frameworks for Engagement; New York: Cambridge University Press, S. 44–75.

Barocas, Solon; Selbst, Andrew D. (2016): Big data's disparate impact, in: California Law Review, 104. Jg., S. 671–732.

Bartlett, Robert P.; Morse, Adair; Stanton, Richard; Wallace, Nancy (2018): Consumer lending discrimination in the FinTech era, UC Berkeley Public Law Research Paper.

Baruh, Lemi (2007): Read at your own risk: shrinkage of privacy and interactive media, in: New Media & Society, 9. Jg., H. 2, S. 187–211.

Baumann, Anne-Luise; Egenberger, Vera; Supik, Linda (2018): Erhebung von Antidiskriminierungsdaten in repräsentativen Wiederholungsbefragungen. Bestandsaufnahme und Entwicklungsmöglichkeiten; Berlin: Antidiskriminierungsstelle des Bundes (ADS).

Beck, Susanne; Grunwald, Armin; Jacob, Kai; Matzner, Tobias (2019): Künstliche Intelligenz und Diskriminierung. Herausforderungen und Lösungsansätze; München: Lernende Systeme - Die Plattform für Künstliche Intelligenz.

Becker, Gary S. (1957/1971): The Economics of Discrimination (Second Edition); Chicago: University of Chicago Press.

Bellamy, Rachel K. E.; Dey, Kuntal; Hind, Michael; Hoffman, Samuel C.; Houde, Stephanie; Kannan, Kalapriya; Lohia, Pranay; Martino, Jacquelyn; Mehta, Sameep; Mojsilovic, Aleksandra (2018): AI Fairness 360: An Extensi-

ble Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias, in: arXiv preprint arXiv:1810.01943.

Berghahn, Sabine; Egenberger, Vera; Klapp, Micha; Klose, Alexander; Liebscher, Doris; Supik, Linda; Tischbirek, Alexander (2016): Evaluation des Allgemeinen Gleichbehandlungsgesetzes, erstellt im Auftrag der Antidiskriminierungsstelle des Bundes; Berlin: Antidiskriminierungsstelle des Bundes, veröffentlicht von Nomos Verlag.

Berghahn, Sabine; Klose, Alexander; Lewalter, Sandra; Liebscher, Doris; Spangenberg, Ulrike; Wersig, Maria (2014): Handbuch „Rechtlicher Diskriminierungsschutz"; Berlin: Antidiskriminierungsstelle des Bundes (ADS).

Berk, Richard; Heidari, Hoda; Jabbari, Shahin; Kearns, Michael; Roth, Aaron (2018): Fairness in criminal justice risk assessments: The State of the Art, in: Sociological Methods & Research, Online first, S. 1–42.

BGH (2014): Urteil des VI. Zivilsenats des Bundesgerichtshofs vom 28.1.2014, VI ZR 156/13 (SCHUFA-Urteil); Karlsruhe: Bundesgerichtshof (BGH).

Bitter, Philip; Uphues, Steffen (2017): Big Data und die Versicherungsgemeinschaft - „Entsolidarisierung" durch Digitalisierung?, ABIDA-Dossier; Münster: Westfälische Wilhelms-Universität Münster, Institut für Informations-, Telekommunikations- und Medienrecht.

Blodgett, Su Lin; Green, Lisa; O'Connor, Brendan (2016): Demographic Dialectal Variation in Social Media: A Case Study of African-American English; in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, S. 1119–1130.

Blodgett, Su Lin; O'Connor, Brendan (2017): Racial disparity in natural language processing: A case study of social media african-american english, in: arXiv preprint arXiv:1707.00061.

Blömeke, Eva; Clement, Michel (2009): Selektives Demarketing - Management von unprofitablen Kunden, in: Schmalenbachs Zeitschrift für betriebswirtschaftliche Forschung, 61. Jg., H. 7, S. 804–835.

Bogen, Miranda; Rieke, Aaron (2018): Help wanted. An Examination of Hiring Algorithms, Equity, and Bias; Washington, D.C., Upturn.

Bolukbasi, Tolga; Chang, Kai-Wei; Zou, James Y.; Saligrama, Venkatesh; Kalai, Adam T. (2016): Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings; in: Advances in Neural Information Processing Systems, S. 4349–4357.

Brantingham, P. Jeffrey; Valasik, Matthew; Mohler, George O. (2018): Does Predictive Policing Lead to Biased Arrests? Results From a Randomized Controlled Trial, in: Statistics and Public Policy, 5. Jg., H. 1, S. 1–6.

Brauneis, Robert; Goodman, Ellen P. (2018): Algorithmic transparency for the smart city, in: Yale Journal of Law and Technology, 20. Jg., S. 103–176.

Breiman, Leo (2001): Statistical modeling: The two cultures, in: Statistical science, 16. Jg., H. 3, S. 199–231.

Brey, Philip (2000): Disclosive Computer Ethics, in: Computer and Society ACM SIGCAS, 30. Jg., H. 4, S. 10–16.

Brey, Philip (2009): Values in Technology and Disclosive Computer Ethics, in: Luciano Floridi (Hrsg.): The Cambridge Handbook of Information and Computer Ethics; Cambridge: Cambridge University Press, S. 41–58.

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

Britz, Gabriele (2008): Einzelfallgerechtigkeit versus Generalisierung. Verfassungsrechtliche Grenzen statistischer Diskriminierung; Tübingen: Mohr Siebeck.

Britz, Gabriele (2010): Informationelle Selbstbestimmung zwischen rechtswissenschaftlicher Grundsatzkritik und Beharren des Bundesverfassungsgerichts, in: Wolfgang Hoffmann-Riem (Hrsg.): Offene Rechtswissenschaft; Tübingen: Mohr Siebeck, S. 561–596.

Brown, Ian; Marsden, Christopher T. (2013): Regulating Code. Good Governance and Better Regulation in the Information Age; Cambridge, MA: MIT Press.

Bruce, Margaret; Adam, Alison (1989): Expert systems and women's lives: a technology assessment, in: Futures, 21. Jg., H. 5, S. 480–497.

Brundage, Miles; Avin, Shahar; Clark, Jack; Toner, Helen; Eckersley, Peter; Garfinkel, Ben; Dafoe, Allan; Scharre, Paul; Zeitzoff, Thomas; Filar, Bobby; Anderson, Hyrum; Roff, Heather; Allen, Gregory C.; Steinhardt, Jacob; Flynn, Carrick; Ó hÉigeartaigh, Seán; Beard, Simon; Belfield, Haydn; Farquhar, Sebastian; Lyle, Clare; Crootof, Rebecca; Evan, Owain; Page, Michael; Bryson, Joanna; Yampolskiy, Roman; Amodei, Dario (2018): The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation; Oxford et al.: Future of Humanity Institute, University of Oxford, Centre for the Study of Existential Risk, University of Cambridge, Center for a New American Security, Electronic Frontier Foundation, OpenAI.

Buchner, Benedikt (2018): DS-GVO Art. 22 Automatisierte Entscheidungen im Einzelfall einschließlich Profiling, in: Jürgen Kühling und Benedikt Buchner (Hrsg.): Datenschutz-Grundverordnung, Bundesdatenschutzgesetz: DS-GVO/BDSG, Kommentar, 2. Aufl.; München: Beck.

Buolamwini, Joy; Gebru, Timnit (2018): Gender shades: Intersectional accuracy disparities in commercial gender classification; in: Conference on Fairness, Accountability and Transparency, Nr. 81, S. 77–91.

Burdon, Mark; Harpur, Paul (2014): Re-conceptualising privacy and discrimination in an age of talent analytics, in: University of New South Wales Law Journal, 37. Jg., H. 2, S. 679–712.

Burrell, Jenna (2016): How the machine 'thinks': Understanding opacity in machine learning algorithms, in: Big Data and Society, 3. Jg., H. 1, S. 1–12.

Busch, Christoph (2018): Algorithmic Accountability, Gutachten im Rahmen des Projekts „Assessing Big Data" (ABIDA); Münster, Karlsruhe: Universität Münster, Karlsruher Institut für Technologie.

BVerfG (1983): Urteil vom 15. Dezember 1983, Az. 1 BvR 209/83 u.a. (Volkszählungsurteil), BVerfGE 65, 1; Karlsruhe: Bundesverfassungsgericht (BVerfG).

BVerfG (1993): Beschluß des Ersten Senats vom 19. Oktober 1993 – 1 BvR 567, 1044/89 (Bürgschaftsverträge), BVerfGE 89, 214; Karlsruhe: Bundesverfassungsgericht (BVerfG).

BVerfG (2018): Beschluss des Ersten Senats vom 11. April 2018 – 1 BvR 3080/09 (Ausstrahlungswirkung des allgemeinen Gleichheitssatzes in das

Zivilrecht), BVerfGE 148, 267–290; Karlsruhe: Bundesverfassungsgericht (BVerfG).

Cabitza, Federico; Rasoini, Raffaele; Gensini, Gian Franco (2017): Unintended consequences of machine learning in medicine, in: JAMA, 318. Jg., H. 6, S. 517–518.

Calders, Toon; Custers, Bart (2013): What Is Data Mining and How Does It Work?, in: Bart Custers, Toon Calders, Bart Schermer und Tal Zarsky (Hrsg.): Discrimination and Privacy in the Information Society; Studies in Applied Philosophy, Epistemology and Rational Ethics, Volume 3; Berlin, Heidelberg: Springer, S. 27–42.

Calders, Toon; Žliobaitė, Indrė (2013): Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures, in: Bart Custers, Toon Calders, Bart Schermer und Tal Zarsky (Hrsg.): Discrimination and Privacy in the Information Society; Studies in Applied Philosophy, Epistemology and Rational Ethics, Volume 3; Berlin, Heidelberg: Springer, S. 43–57.

Caliskan, Aylin; Bryson, Joanna J.; Narayanan, Arvind (2017): Semantics derived automatically from language corpora contain human-like biases, in: Science, 356. Jg., H. 6334, S. 183–186.

Caruana, Rich; Lou, Yin; Gehrke, Johannes; Koch, Paul; Sturm, Marc; Elhadad, Noemie (2015): Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission; in: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, S. 1721–1730, veröffentlicht von ACM.

Castelluccia, Claude; Le Métayer, Daniel (2019): Understanding algorithmic decision-making: Opportunities and challenges. Study for Panel for the Future of Science and Technology; Brussels: European Parliament, European Parliamentary Research Service, Scientific Foresight Unit (STOA).

Castelvecchi, Davide (2016): The Black Box of AI, in: Nature, 538. Jg., H. October 6, 2016, S. 20–23.

Cate, Fred H.; Mayer-Schönberger, Viktor (2013): Notice and consent in a world of Big Data, in: International Data Privacy Law, 3. Jg., H. 2, S. 67–73.

Chamorro-Premuzic, Tomas; Akhtar, Reece; Winsborough, Dave; Sherman, Ryne A. (2017): The datafication of talent: how technology is advancing the science of human potential at work, in: Current Opinion in Behavioral Sciences, 18. Jg., H. Dec., S. 13–16.

Chamorro-Premuzic, Tomas; Winsborough, Dave; Sherman, Ryne A.; Hogan, Robert (2016): New Talent Signals: Shiny New Objects or a Brave New World?, in: Industrial and Organizational Psychology, 9. Jg., H. 3, S. 621–640.

Chen, Le; Ma, Ruijun; Hannák, Anikó; Wilson, Christo (2018): Investigating the Impact of Gender on Rank in Resume Search Engines; in: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, S. 651, veröffentlicht von ACM.

Chen, Min; Mao, Shiwen; Liu, Yunhao (2014): Big Data: A Survey, in: Mobile Networks and Applications, 19. Jg., H. 2, S. 171–209.

Chouldechova, Alexandra (2017): Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, in: Big data, 5. Jg., H. 2, S. 153–163.

Chouldechova, Alexandra; Benavides-Prado, Diana; Fialko, Oleksandr; Vaithianathan, Rhema (2018): A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions, Proceedings of the 1st Conference on Fairness, Accountability and Transparency; Proceedings of Machine Learning Research: PMLR.

Christl, Wolfie (2017): Corporate Surveillance in Everyday Life. How Companies Collect, Combine, Analyze, Trade, and Use Personal Data on Billions; Vienna: Cracked Labs.

Christl, Wolfie; Spiekermann, Sarah (2016): Networks of Control. A Report on Corporate Surveillance, Digital Tracking, Big Data & Privacy; Wien: Facultas Verlags- und Buchhandels AG.

Citron, Danielle Keats (2016): (Un)Fairness of Risk Scores in Criminal Sentencing; in: Forbes, Onlineartikel vom 13.7.2016, abrufbar unter: https://www.forbes.com/sites/daniellecitron/2016/07/13/unfairness-of-risk-scores-in-criminal-sentencing/#4774e7c24479 (zuletzt abgerufen am 28.8.2019).

Citron, Danielle Keats; Pasquale, Frank (2014): The scored society: due process for automated predictions, in: Washington Law Review, 89. Jg., H. 1, S. 101–133.

Conger, Kate; Fausset, Richard; Kovaleski, Serge F. (2019): San Francisco Bans Facial Recognition Technology, in: New York Times, Onlineartikel vom 14.05.2019, abrufbar unter: https://www.nytimes.com/2019/05/14/us/facial-recognition-ban-san-francisco.html (letzter Zugriff am 28.8.2019).

Constantiou, Ioanna D.; Kallinikos, Jannis (2015): New games, new rules: big data and the changing context of strategy, in: Journal of Information Technology, 30. Jg., H. 1, S. 44–57.

Cormen, Thomas H.; Leiserson, Charles E.; Rivest, Ronald; Stein, Clifford; Molitor, Paul (2010): Algorithmen - Eine Einführung, 3. Aufl.; München: Oldenbourg Verlag.

Council of Europe (2019): Unboxing Artificial Intelligence: 10 steps to protect Human Rights; Straßbourg: Council of Europe, Commissioner for Human Rights.

Courtland, Rachel (2018): Bias detectives: the researchers striving to make algorithms fair. As machine learning infiltrates society, scientists are trying to help ward off injustice, in: Nature, 558. Jg., H. June 20, 2018, S. 357–360.

Crawford, Kate (2013): The Hidden Biases in Big Data, in: Harvard Business Review, Onlineartikel vom 1.4.2014, abrufbar unter: https://hbr.org/2013/04/the-hidden-biases-in-big-data?autocomplete=true (zuletzt abgerufen am 29.8.2019).

Crawford, Kate; Whittaker, Meredith; Elish, Madeleine Clare; Barocas, Solon; Plasek, Aaron; Ferryman, Kadija (2016): The AI Now Report. The Social and Economic Implications of Artificial Intelligence, Technologies in the Near-Term. A summary of the AI Now public symposium, hosted by the White House and New York University's Information Law Institute, July 7th, 2016; New York: New York University, AI Now Institute.

Cummings, Mary (2004a): Automation bias in intelligent time critical decision support systems; in: AIAA 1st Intelligent Systems Technical Conference, S. Paper AIAA-6113.

Cummings, Mary L. (2004b): Creating moral buffers in weapon control interface design, in: IEEE Technology and Society Magazine, 23. Jg., H. 3, S. 28–33.

Custers, Bart (2013): Data Dilemmas in the Information Society: Introduction and Overview, in: Bart Custers, Toon Calders, Bart Schermer und Tal Zarsky (Hrsg.): Discrimination and Privacy in the Information Society; Studies in Applied Philosophy, Epistemology and Rational Ethics, Volume 3; Berlin, Heidelberg: Springer, S. 3–26.

Das, Sauvik; Kramer, Adam D. I. (2013): Self-Censorship on Facebook; in: Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (ICWSM), S. 120–127, veröffentlicht von Association for the Advancement of Artificial Intelligence.

Dastin, Jeffrey (2018): Amazon scraps secret AI recruiting tool that showed bias against women, in: Reuters Business News, Onlineartikel veröffentlicht von Reuters vom 2018-10-10, abrufbar unter: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G (letzter Zugriff am 28.8.2019).

Datta, Amit; Datta, Anupam; Makagon, Jael; Mulligan, Deirdre. K.; Tschantz, Michael Carl (2018): Discrimination in Online Personalization: A Multidisciplinary Inquiry, in: Proceedings of Machine Learning Research, 81. Jg., S. 1–15.

Datta, Amit; Tschantz, Michael Carl; Datta, Anupam (2015): Automated experiments on ad privacy settings, in: Proceedings on Privacy Enhancing Technologies, 2015. Jg., H. 1, S. 92–112.

Däubler, Wolfgang (2018): AGG §1 Ziel des Gesetzes, in: Wolfgang Däubler und Martin Bertzbach (Hrsg.): Allgemeines Gleichbehandlungsgesetz – Handkommentar, 4. Aufl.; Baden-Baden: Nomos.

De-Arteaga, Maria; Romanov, Alexey; Wallach, Hanna; Chayes, Jennifer; Borgs, Christian; Chouldechova, Alexandra; Geyik, Sahin; Kenthapadi, Krishnaram; Kalai, Adam Tauman (2019): Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting; in: Proceedings of the Conference on Fairness, Accountability, and Transparency, S. 120–128, veröffentlicht von ACM.

de Laat, Paul B. (2017): Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability?, in: Philosophy & Technology, S. 1–17.

DerStandard (2019): Volksanwaltschaft kritisiert AMS-Algorithmus, in: DerStandard, Onlineartikel vom 10.03.2019, abrufbar unter: https://apps.derstandard.de/privacywall/story/2000099270837/volksanwaltschaft-kritisiert-ams-algorithmus-in-der-krikik-der (letzter Zugriff am 28.8.2019).

Desai, Deven R.; Kroll, Joshua A. (2017): Trust but Verify: A Guide to Algorithms and the Law, in: Harvard Journal of Law & Technology, 31. Jg., H. 1, S. 1.

Deutscher Ethikrat (2017): Big Data und Gesundheit – Datensouveränität als informationelle Freiheitsgestaltung, Stellungnahme; Berlin: Deutscher Ethikrat.

Dewenter, Ralf; Lüth, Hendrik (2018): Datenhandel und Plattformen, Gutachten im Rahmen des Projekts „Assessing Big Data“ (ABIDA); Münster, Karlsruhe: Universität Münster, Karlsruher Institut für Technologie.

Diakopoulos, Nicholas (2014): Algorithmic Accountability Reporting: On the Investigation of Black Boxes; New York: Columbia University Academic Commons.

Dieterich, William; Mendoza, Christina; Brennan, Tim (2016): COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity. Performance of the COMPAS Risk Scales in Broward County. Technical report, July 2016: Northpointe Inc.

Dixon, Pam; Gellman, Robert (2014): The Scoring of America: How Secret Consumer Scores Threaten Your Privacy and Your Future; Washington, D.C.: World Privacy Forum.

Domingos, Pedro (2012): A Few Useful Things to Know About Machine Learning, in: Communications of the ACM, 55. Jg., H. 10, S. 78–87.

Dorfleitner, Gregor; Hornuf, Lars (2018): Neue digitale Akteure und ihre Rolle in der Finanzwirtschaft. Eine Analyse des deutschen Marktes unter besonderer Berücksichtigung von Datenschutzaspekten, Gutachten im

Rahmen des Projekts „Assessing Big Data“ (ABIDA); Münster, Karlsruhe: Universität Münster, Karlsruher Institut für Technologie.

Dosilovic, F. K.; Brcic, M.; Hlupic, N. (2018): Explainable artificial intelligence: A survey; in: 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2018 - Proceedings, S. 210–215.

Dwork, Cynthia; Mulligan, Deirdre K. (2013): It’s Not Privacy, And It’s Not Fair, in: Stanford Law Review Online, 66. Jg., S. 35–40.

Dzida, Boris (2017): Big Data im Arbeitsrecht, in: Neue Zeitschrift für Arbeitsrecht (NZA), 34. Jg., H. 9, S. 541–546.

Dzida, Boris; Groh, Naemi (2018): Diskriminierung nach dem AGG beim Einsatz von Algorithmen im Bewerbungsverfahren, in: Neue Juristische Wochenschrift (NJW), 71. Jg., H. 27, S. 1917–1922.

Ebert, Ina (2019): Allgemeines Gleichbehandlungsgesetz (AGG), in: Reiner Schulze (Hrsg.): Bürgerliches Gesetzbuch - Handkommentar, 10. Aufl.; Baden-Baden: Nomos.

Eckhouse, Laurel; Lum, Kristian; Conti-Cook, Cynthia; Ciccolini, Julie (2019): Layers of Bias: A Unified Approach for Understanding Problems With Risk Assessment, in: Criminal Justice and Behavior, 46. Jg., H. 2, S. 185–209.

Edelman, Benjamin G.; Luca, Michael (2014): Digital discrimination: The case of Airbnb.com, Harvard Business School NOM Unit Working Paper, No. 14-054; Boston: Harvard Business School.

Edelman, Benjamin; Luca, Michael; Svirsky, Dan (2017): Racial discrimination in the sharing economy: Evidence from a field experiment, in: American Economic Journal: Applied Economics, 9. Jg., H. 2, S. 1–22.

EDPS (2017): Opinion 4/2017 on the Proposal for a Directive on certain aspects concerning contracts for the supply of digital content; Brussels: European Data Protection Supervisor (EDPS).

EDPS (2018): Guidelines on the protection of personal data in IT governance and IT management of EU institutions; Brussels: European Data Protection Supervisor (EDPS).

Edwards, Lilian; Veale, Michael (2017): Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for, in: Duke Law & Technology Review, 16. Jg., H. 1, S. 18–84.

Ehsan, Upol; Tambwekar, Pradyumna; Chan, Larry; Harrison, Brent; Riedl, Mark (2019): Automated rationale generation: a technique for explainable AI and its effects on human perceptions, in: arXiv preprint arXiv:1901.03729.

Epp, Clayton; Lippold, Michael; Mandryk, Regan L. (2011): Identifying emotional states using keystroke dynamics; in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, S. 715–724, veröffentlicht von ACM.

Ernst, Christian (2017): Algorithmische Entscheidungsfindung und personenbezogene Daten, in: Juristenzeitung, 72. Jg., H. 21, S. 1026–1036.

Eschholz, Stefanie (2017): Big Data-Scoring unter dem Einfluss der Datenschutz-Grundverordnung, in: Datenschutz und Datensicherheit – DuD, 41. Jg., H. 3, S. 180–185.

Eubanks, Virginia (2017): Automating Inequality. How High-Tech Tools Profile, Police, and Punish the Poor; New York: St. Martin's Press.

Ezrachi, Ariel; Stucke, Maurice E. (2016): Virtual Competition. The Promise and Perils of the Algorithm-Driven Economy; Cambridge, London: Harvard University Press.

Fang, Hanming; Moro, Andrea (2011): Theories of statistical discrimination and affirmative action: A survey, in: Jess Benhabib, Matthew O. Jackson und Alberto Bisin (Hrsg.): Handbook of Social Economics, Volume 1A; Amsterdam: North-Holland (Elsevier), S. 133–200.

Fanta, Alexander (2018): Österreichs Jobcenter richten künftig mit Hilfe von Software über Arbeitslose, in: Netzpolitik.org, Onlineartikel vom 18.10.2018, abrufbar unter: https://netzpolitik.org/2018/oesterreichs-jobcenter-richten-kuenftig-mit-hilfe-von-software-ueber-arbeitslose/ (letzter Zugriff am 28.8.2019.

Favaretto, Maddalena; De Clercq, Eva; Elger, Bernice Simone (2019): Big Data and discrimination: perils, promises and solutions. A systematic review, in: Journal of Big Data, 6. Jg., H. 1, S. 1–27.

Ferguson, Andrew Guthrie (2017): The Truth About Predictive Policing and Race, in: The Appeal, Onlineartikel vom 07.12.2017, abrufbar unter: https://theappeal.org/the-truth-about-predictive-policing-and-race-b87cf7c070b1/ (letzter Zugriff am 28.8.2019).

Ferretti, Federico (2017): Not-So-Big and Big Credit Data Between Traditional Consumer Finance, FinTechs, and the Banking Union: Old and New Challenges in an Enduring EU Policy and Legal Conundrum, in: Global Jurist, 18. Jg., H. 1, 1–41.

Fink, Katherine (2018): Opening the government's black boxes: freedom of information and algorithmic accountability, in: Information, Communication & Society, 21. Jg., H. 10, S. 1453–1471.

Forbrukerrådet (2018): Deceived by design. How tech companies use dark patterns to discourage us from exercising our rights to privacy; Oslo: Forbrukerrådet.

FRA (2018): #BigData: Discrimination in data-supported decision making, FRA Focus; Vienna: European Union Agency for Fundamental Rights (FRA).

FRA (2019): Data quality and artificial intelligence - mitigating bias and error to protect fundamental rights, FRA Focus; Vienna: European Union Agency for Fundamental Rights (FRA).

Franke, Bernhard; Schlichtmann, Gisbert (2018): AGG §20 Zulässige unterschiedliche Behandlung, in: Wolfgang Däubler und Martin Bertzbach (Hrsg.): Allgemeines Gleichbehandlungsgesetz – Handkommentar, 4. Aufl.; Baden-Baden: Nomos.

Freeman, Katherine (2016): Algorithmic Injustice: How the Wisconsin Supreme Court Failed to Protect Due Process Rights in State v. Loomis, in: North Carolina Journal of Law & Technology, 18. Jg., H. 5, S. 75–106.

Friedler, Sorelle A.; Scheidegger, Carlos; Venkatasubramanian, Suresh; Choudhary, Sonam; Hamilton, Evan P.; Roth, Derek (2019): A comparative study of fairness-enhancing interventions in machine learning; in: FAT*

‘19 Proceedings of the Conference on Fairness, Accountability, and Transparency, January 29 – 31, 2019, Atlanta, GA, USA, S. 329–338, veröffentlicht von ACM.

Friedman, Batya; Nissenbaum, Helen (1996): Bias in Computer Systems, in: ACM Transactions on Information Systems, 14. Jg., H. 3, S. 330–347.

Fröhlich, Wiebke; Spiecker genannt Döhmann, Indra (2018): Können Algorithmen diskriminieren?, in: Verfassungsblog (VerfBlog), Onlineartikel vom 26.12.2018, abrufbar unter: https://verfassungsblog.de/koennen-algorithmen-diskriminieren/ (letzter Zugriff am 27.8.2019).

FTC (2016): Big Data. A Tool for Inclusion or Exclusion?; Washington, D.C.: Federal Trade Commission (FTC).

Fu, S.; He, H.; Hou, Z. (2014): Learning Race from Face: A Survey, in: IEEE Transactions on Pattern Analysis and Machine Intelligence, 36. Jg., H. 12, S. 2483-2509.

Galhotra, Sainyam; Brun, Yuriy; Meliou, Alexandra (2017): Fairness testing: testing software for discrimination; in: Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering, S. 498–510, veröffentlicht von ACM.

Gandy Jr., Oscar H. (2009): Coming to Terms with Chance: Engaging Rational Discrimination and Cumulative Disadvantage; Farnham, Burlington: Ashgate.

Gandy Jr., Oscar H. (2010): Engaging rational discrimination: exploring reasons for placing regulatory constraints on decision support systems, in: Ethics and Information Technology, 12. Jg., H. 1, S. 1–14.

Garg, Nikhil; Schiebinger, Londa; Jurafsky, Dan; Zou, James (2018): Word embeddings quantify 100 years of gender and ethnic stereotypes, in: Proceedings of the National Academy of Sciences, 115. Jg., H. 16, S. E3635–E3644.

Garvie, Clare; Bedoya, Alvaro M.; Frankle, Jonathan (2016): The perpetual line-up: Unregulated police face recognition in America; Washington, D.C.: Georgetown Law, Center on Privacy & Technology.

Géron, Aurélien (2018): Praxiseinstieg Machine Learning mit Scikit-Learn und TensorFlow: Konzepte, Tools und Techniken für intelligente Systeme (übersetzt von Kristian Rother); Heidelberg: O'Reilly.

Gilheany, John; Wang, David; Xi, Stephen (2015): The model minority? Not on Airbnb.com: A hedonic pricing model to quantify racial bias against Asian Americans, in: Technology Science, Onlineartikel vom 01.09.2015, abrufbar unter: https://techscience.org/a/2015090104/ (letzter Zugriff am 28.8.2019).

Gillum, Jack; Tobin, Ariana (2019): Facebook Won't Let Employers, Landlords or Lenders Discriminate in Ads Anymore, in: ProPublica, Onlineartikel vom 19.03.2019, abrufbar unter: https://www.propublica.org/article/facebook-ads-discrimination-settlement-housing-employment-credit (letzter Zugriff am 28.8.2019).

Goldfarb, Avi; Tucker, Catherine (2017): Digital Economics, NBER Working Paper 23684; Cambridge, MA: National Bureau of Economic Research.

Goodman, Bryce; Flaxman, Seth (2017): European Union regulations on algorithmic decision-making and a "right to explanation", in: AI Magazine, 38. Jg., H. 3, S. 50–57.

Goodman, Bryce W. (2016): Economic Models of (Algorithmic) Discrimination; in: Machine Learning and the Law, NIPS Symposium, 8 December, 2016, Barcelona, Spain.

Grimmelmann, James (2005): Regulation by Software, in: Yale Law Journal, 114. Jg., H. 7, S. 1721–1758.

Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; Pedreschi, D. (2018): A survey of methods for explaining black box models, in: ACM Computing Surveys, 51. Jg., H. 5., S. 1–42.

Gurovich, Yaron; Hanani, Yair; Bar, Omri; Nadav, Guy; Fleischer, Nicole; Gelbman, Dekel; Basel-Salmon, Lina; Krawitz, Peter M.; Kamphausen, Susanne B.; Zenker, Martin (2019): Identifying facial phenotypes of genetic disorders using deep learning, in: Nature medicine, 25. Jg., H. 1, S. 60.

Hacker, Philipp; Petkova, Bilyana (2017): Reining in the Big Promise of Big Data: Transparency, Inequality, and New Regulatory Frontiers, in: Northwestern Journal of Technology and Intellectual Property, 15. Jg., H. 1.

Hand, David J. (2006): Classifier Technology and the Illusion of Progress, in: Statistical Science, 21. Jg., H. 1, S. 1–14.

Hannák, Anikó.; Soeller, Gary; Lazer, David; Mislove, Alan; Wilson, Christo (2014): Measuring price discrimination and steering on e-commerce web sites; in: Proceedings of the 2014 conference on internet measurement conference, S. 305–318, veröffentlicht von ACM.

Hannák, Anikó.; Wagner, Claudia; Garcia, David; Mislove, Alan; Strohmaier, Markus; Wilson, Christo (2017): Bias in online freelance marketplaces: Evidence from taskrabbit and fiverr; in: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, S. 1914–1933, veröffentlicht von ACM.

Hannák, Anikó.; Wagner, Claudia; Garcia, David; Strohmaier, Markus; Wilson, Christo (2016): Bias in online freelance marketplaces: Evidence from taskrabbit; in: Proceedings DAT Workshop, S. 1914–1933, veröffentlicht von ACM.

Hänold, Stefanie (2018): Profiling and Automated Decision-Making: Legal Implications and Shortcomings, in: Ugo Pagallo, Marcelo Corrales, Mark Fenwick und Nikolaus Forgó (Hrsg.): Robotics, AI and the Future of Law; Singapore: Springer Singapore, S. 123–153.

Hänold, Stefanie (2019): Profiling und automatisierte Einzelentscheidungen im Versicherungsbereich. Bericht im Rahmen des Projekts „Assessing Big Data“ (ABIDA); Hannover: Institutionelles Repositorium der Leibniz Universität Hannover.

Hargittai, Eszter (2015): Is bigger always better? Potential biases of big data derived from social network sites, in: The ANNALS of the American Academy of Political and Social Science, 659. Jg., H. 1, S. 63–76.

Härtel, Ines (2019): Digitalisierung im Lichte des Verfassungsrechts – Algorithmen, Predictive Policing, autonomes Fahren, in: Landes- und Kommunalverwaltung, 29. Jg., H. 2, S. 49–60.

Helberger, Natali (2016): Profiling and Targeting Consumers in the Internet of Things - A new Challenge for Consumer Law, in: Reiner Schulze und Dirk Staudenmayer (Hrsg.): Digital Revolution: Challenges for Contract Law in Practice; Baden-Baden: Nomos, S. 135–165.

Hellman, Deborah (1998): Two Types of Discrimination: The Familiar and the Forgotten, in: California Law Review, 86. Jg., H. 2, S. 315–361.

Hellman, Deborah (2008): When is Discrimination Wrong?; Cambridge, London: Harvard University Press.

Hildebrandt, Mireille (2018): Algorithmic regulation and the rule of law, in: Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376. Jg., H. 2128, S. 20170355.

Hildebrandt, Mireille; Gutwirth, Serge (Hrsg.) (2008): Profiling the European Citizen. Cross-Disciplinary Perspectives; Dordrecht: Springer.

Hill, Robin K. (2016): What an algorithm is, in: Philosophy & Technology, 29. Jg., H. 1, S. 35–59.

Hinz, Thomas; Ausprung, Katrin (2017): Diskriminierung auf dem Wohnungsmarkt, in: Albert Scherr, Aladin El-Mafaalani und Gökçen Yüksel (Hrsg.): Handbuch Diskriminierung; Wiesbaden: Springer VS, S. 387–406.

Hoeren, Thomas; Kolany-Raiser, Barbara (Hrsg.) (2018): Big Data in Context. Legal, Social and Technological Insights, Springer Briefs in Law; Cham: Springer Open.

Hoeren, Thomas; Niehoff, Maurice (2018): KI und Datenschutz - Begründungserfordernisse automatisierter Entscheidungen, in: Rechtswissenschaft (RW) - Zeitschrift für rechtswissenschaftliche Forschung, 9. Jg., H. 1, S. 47–66.

Hoffmann-Riem, Wolfgang (1998): Informationelle Selbstbestimmung in der Informationsgesellschaft - Auf dem Wege zu einem neuen Konzept des Datenschutzes, in: Archiv des Öffentlichen Rechts, 123. Jg., H. 4, S. 513–540.

Hoffmann-Riem, Wolfgang (2017): Verhaltenssteuerung durch Algorithmen - Eine Herausforderung für das Recht, in: Archiv des öffentlichen Rechts, 142. Jg., H. 1, S. 1–42.

Holl, Jürgen; Kernbeiß, Günter; Wagner-Pinter, Michael (2018): Das AMS-Arbeitsmarktchancen-Modell. Dokumentation zur Methode; Wien: Synthesis Forschung.

Hurley, Mikella; Adebayo, Julius (2016): Credit Scoring in the Era of Big Data, in: Yale Journal of Law & Technology, 18. Jg., H. 1, S. 148–216.

Illinois Attorney General (2017): Madigan Probes National Job Search Sites Over Potential Age Discrimination. Attorney General Madigan Calls on Career Search Sites to Explain Potential Age Discrimination Violations Against Older Job Seekers, Onlineartikel vom 02.03.2017, abrufbar unter: http://www.illinoisattorneygeneral.gov/pressroom/2017_03/20170302.html (letzter Zugriff am 28.8.2019).

Ingold, David; Soper, Spencer (2016): Amazon Doesn't Consider the Race of Its Customers. Should It?, in: Bloomberg, Onlineartikel vom 21.04.2016, update 01.05.2016, abrufbar unter: https://www.bloomberg.com/graphics/2016-amazon-same-day/ (letzter Zugriff am 28.8.2019).

Isaac, William; Lum, Kristian (2018): Setting the Record Straight on Predictive Policing and Race, Onlineartikel veröffentlicht von The Appeal vom 03.01.2018, abrufbar unter: https://theappeal.org/setting-the-record-straight-on-predictive-policing-and-race-fe588b457ca2/ (letzter Zugriff am 28.8.2019).

Jernigan, Carter; Mistree, Behram F.T. (2009): Gaydar: Facebook friendships expose sexual orientation, in: First Monday, 14. Jg., H. 10.

Jordan, M. I.; Mitchell, T. M. (2015): Machine learning: Trends, perspectives, and prospects, in: Science, 349. Jg., H. 6245, S. 255–260.

Just, Natascha; Latzer, Michael (2016): Governance by Algorithms: Reality Construction by Algorithmic Selection on the Internet, in: Media, Culture & Society, 39. Jg., H. 2, S. 238–258.

Kallinikos, Jannis (2011): Governing through Technology: Information Artefacts and Social Practice; Basingstoke: Palgrave Macmillan.

Kamp, Meike; Rost, Martin (2013): Kritik an der Einwilligung. Ein Zwischenruf zu einer fiktiven Rechtsgrundlage in asymmetrischen Machtverhältnissen, in: Datenschutz und Datensicherheit, 37. Jg., H. 2, S. 80–84.

Kamp, Meike; Weichert, Thilo (2005): Scoringsysteme zur Beurteilung der Kreditwürdigkeit - Chancen und Risiken für Verbraucher; Kiel: Unabhängigen Landeszentrum für Datenschutz Schleswig-Holstein (ULD).

Kant, Immanuel (1786/1977): Grundlegung zur Metaphysik der Sitten. In: Immanuel Kant: Werke in zwölf Bänden. Band 7; Erstdruck: Riga (Hartknoch) 1785. Der Text folgt der 2. (verbesserten) Auflage, Riga (Hartknoch) 1786.; Frankfurt am Main: Gemeinfreie Ausgabe über Zeno.org, http://www.zeno.org/nid/20009189599 (zuletzt abgerufen am 29.8.2019).

Kasperkevic, Jana (2015): Google says sorry for racist auto-tag in photo app, in: The Guardian, Onlineartikel vom 01.07.2015, abrufbar unter: https://www.theguardian.com/technology/2015/jul/01/google-sorry-racist-auto-tag-photo-app (letzter Zugriff am 28.8.2019).

Kay, Matthew; Matuszek, Cynthia; Munson, Sean A. (2015): Unequal representation and gender stereotypes in image search results for occupations; in: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, S. 3819–3828, veröffentlicht von ACM.

Kettner, Elisa; Thorun, Christian; Kleinhans, Jan-Peter (2018): Big Data im Bereich Heim und Freizeit. Smart Living: Status Quo und Entwicklungstendenzen, Gutachten im Rahmen des Projekts „Assessing Big Data“ (ABIDA); Münster, Karlsruhe: Universität Münster, Karlsruher Institut für Technologie.

Kim, Pauline T. (2016): Data-driven discrimination at work, in: William & Mary Law Review, 58. Jg., H. 3, S. 857–936.

Kiritchenko, Svetlana; Mohammad, Saif M. (2018): Examining gender and race bias in two hundred sentiment analysis systems, in: arXiv preprint arXiv:1805.04508.

Kitchin, Rob (2017): Thinking critically about and researching algorithms, in: Information, Communication & Society, 20. Jg., H. 1, S. 14–29.

Klare, Brendan F.; Burge, Mark J.; Klontz, Joshua C.; Bruegge, Richard W. Vorder; Jain, Anil K. (2012): Face recognition performance: Role of demographic information, in: IEEE Transactions on Information Forensics and Security, 7. Jg., H. 6, S. 1789–1801.

Klebert, Florian; Shirazi, Fatemeh; Simo, Hervais; Wüchner, Tobias; Buchmann, Johannes; Pretschner, Alexander; Waidner, Michael (2012): State of Online Privacy: A Technical Perspective, in: Johannes Buchmann (Hrsg.): Internet Privacy. Eine multidisziplinäre Bestandsaufnahme / A multidisciplinary analysis (acatech STUDIE); Heidelberg u.a.: Springer Verlag, S. 189–279.

Kleinberg, Jon; Ludwig, Jens; Mullainathan, Sendhil; Sunstein, Cass R. (2019): Discrimination in the Age of Algorithms; Cambridge, MA: National Bureau of Economic Research.

Klingele, Cecelia (2015): The promises and perils of evidence-based corrections, in: Notre Dame Law Review, 91. Jg., H. 2, S. 101–151.

Kolany-Raiser, Barbara; Heil, Reinhard; Orwat, Carsten; Hoeren, Thomas (Hrsg.) (2018): Big Data und Gesellschaft. Eine multidisziplinäre Annäherung; Wiesbaden: Springer VS.

Kornwachs, Klaus (2018): Arbeit 4.0 – People Analytics. Führungsinformationssysteme: Soziologische, psychologische, wissenschaftsphilosophisch-ethische Überlegungen zum Einsatz von Big Data in Personalmanagement und Personalführung, Gutachten im Rahmen des Projekts „Assessing Big Data“ (ABIDA); Argenbühl-Eglofs: Büro für Kultur und Technik.

Kosinski, Michal; Stillwell, David; Graepel, Thore (2013): Private traits and attributes are predictable from digital records of human behavior, in: Proceedings of the National Academy of Sciences, 110. Jg., H. 15, S. 5802–5805.

Kroll, Joshua A. (2018): The fallacy of inscrutability, in: Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376. Jg., H. 2133, S. 1–14.

Lambrecht, Anja; Tucker, Catherine E. (2019): Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads, in: Management Science (Articles in Adance).

Lang, Caroline; Barton, Hannah (2015): Just untag it: Exploring the management of undesirable Facebook photos, in: Computers in Human Behavior, 43. Jg., H. (Feb.), S. 147–155.

Lecuyer, Mathias; Spahn, Riley; Spiliopolous, Yannis; Chaintreau, Augustin; Geambasu, Roxana; Hsu, Daniel (2015): Sunlight: Fine-grained targeting detection at scale with statistical confidence; in: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, S. 554–566, veröffentlicht von ACM.

Lehr, David; Ohm, Paul (2017): Playing with the Data: What Legal Scholars Should Learn About Machine Learning, in: UC Davis Law Review, 51. Jg., S. 653–717.

Leis, Miriam; Petzka, Henning; Rüping, Stefan; Voss, Angelika (2018): Maschinelles Lernen - Einordnung, Konzepte, Methoden und Grenzen, in: Inga Döbel, Miriam Leis, Manuel Molina Vogelsang, Dmitry Neustroev, Henning Petzka, Stefan Rüping, Angelika Voss, Martin Wegele und Juliane Welz (Hrsg.): Maschinelles Lernen - Kompetenzen, Anwendungen und Forschungsbedarf: Fraunhofer IAIS, Fraunhofer IMW, Fraunhofer Zentrale, S. 7–52.

Lerman, Jonas (2013): Big data and its exclusions, in: Stanford Law Review Online, 66. Jg., H. Sep., S. 55–63.

Lessig, Lawrence (1999): Code and other laws of cyberspace; New York: Basic Books.

Lessig, Lawrence (2006): Code Version 2.0; New York: Basic Books.

Levy, Karen; Barocas, Solon (2017): Designing against discrimination in online markets, in: Berkeley Technolgy Law Journal, 32. Jg., H. 3, S. 1183–1237.

Linoff, Gordon S.; Berry, Michael J. A. (2011): Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management; Indianapolis: Wiley.

Lippert-Rasmussen, Kasper (2007): Nothing personal: On statistical discrimination, in: Journal of Political Philosophy, 15. Jg., H. 4, S. 385–403.

Lischka, Konrad; Klingel, Anita (2017): Wenn Maschinen Menschen bewerten. Internationale Fallbeispiele für Prozesse algorithmischer Entscheidungsfindung - Arbeitspapier -; Gütersloh: Bertelsmann Stiftung.

Lorenz, Wilhelm (1993): Diskriminierung, in: Bernd-Thomas Ramb und Manfred Tietzel (Hrsg.): Ökonomische Verhaltenstheorie; München: Vahlen, S. 119–147.

Lum, Kristian; Isaac, William (2016): To predict and serve?, in: Significance, 13. Jg., H. 5, S. 14–19.

Mantelero, Alessandro (2016): Personal data for decisional purposes in the age of analytics: From an individual to a collective dimension of data protection, in: Computer Law and Security Review, 32. Jg., H. 2, S. 238–255.

Marder, Ben; Joinson, Adam; Shankar, Avi; Houghton, David (2016): The extended 'chilling' effect of Facebook: The cold reality of ubiquitous social networking, in: Computers in Human Behavior, 60. Jg., H. Jul., S. 582–592.

Marler, Janet H.; Boudreau, John W. (2017): An evidence-based review of HR Analytics, in: The International Journal of Human Resource Management, 28. Jg., H. 1, S. 3–26.

Marthews, Alex; Tucker, Catherine E. (2017): Government Surveillance and Internet Search Behavior; Cambridge, MA: Digital Fourth and MIT Sloan School of Management.

Martin, Kirsten (2013): Transaction costs, privacy, and trust: The laudable goals and ultimate failure of notice and choice to respect privacy online, in: First Monday, 18. Jg., H. 12, Onlineartikel vom 2.12.2013, abrufbar unter: https://firstmonday.org/ojs/index.php/fm/article/view/4838/3802 (zuletzt abgerufen am 27.8.2019).

Martini, Mario (2017): Algorithmen als Herausforderung für die Rechtsordnung, in: Juristenzeitung, 72. Jg., H. 21, S. 1017–1025.

Martini, Mario (2018): DS-GVO Art. 22 Automatisierte Entscheidungen im Einzelfall einschließlich Profiling, in: Boris P. Paal und Daniel A. Pauly (Hrsg.): Datenschutz-Grundverordnung Bundesdatenschutzgesetz DS-GVO BDSG, Kommentar, 2. Aufl.; München: Beck.

Martini, Mario; Nink, David (2017): Wenn Maschinen entscheiden … – vollautomatisierte Verwaltungsverfahren und der Persönlichkeitsschutz, in: Neue Zeitschrift für Verwaltungsrecht - Extra, 36. Jg., H. 10, S. 1–14.

Mathur, Arunesh; Acar, Gunes; Friedman, Michael; Lucherini, Elena; Mayer, Jonathan; Chetty, Marshini; Narayanan, Arvind (2019): Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites; Princeton: Princeton University.

Matsakis, Louise (2019): Facebook's Ad System Might Be Hard-Coded for Discrimination, in: Wired, Onlineartikel vom 06.04.2019, abrufbar unter: https://www.wired.com/story/facebooks-ad-system-discrimination/ (letzter Zugriff am 28.8.2019).

Matz, S. C.; Kosinski, M.; Nave, G.; Stillwell, D. J. (2017): Psychological targeting as an effective approach to digital mass persuasion, in: Proceedings of the National Academy of Sciences (PNAS), 114. Jg., H. 48, S. 12714–12719.

Matz, Sandra C.; Netzer, Oded (2017): Using Big Data as a window into consumers' psychology, in: Current Opinion in Behavioral Sciences, 18. Jg., H. Dec., S. 7–12.

Matzat, Lorenz; Zielinski, Lukas; Cocco, Miriam; Penner, Kristina; Spielkamp, Matthias; Gießler, Sebastian; Lang, Sebastian; Thiel, Veronika (2019): Atlas der Automatisierung: Automatisierung und Teilhabe in Deutschland; Berlin: AW AlgorithmWatch gGmbH.

Mayer, Jonathan; Mutchler, Patrick; Mitchell, John C. (2016): Evaluating the privacy properties of telephone metadata, in: Proceedings of the National Academy of Sciences, 113. Jg., H. 20, S. 5536–5541.

McDonald, Aleecia M.; Cranor, Lorrie Faith (2008): The cost of reading privacy policies, in: I/S: A Journal of Law and Policy for the Information Society, 4. Jg., H. 3, S. 543–568.

Medina, E. (2015): Rethinking algorithmic regulation, in: Kybernetes, 44. Jg., H. 6-7, S. 1005-1019.

Merz, Christina (2016): Predictive Policing - Polizeiliche Strafverfolgung in Zeiten von Big Data; Dossier des Projekts „Assessing Big Data" (ABIDA); Karlsruhe: Institut für Technikfolgenabschätzung und Systemanalyse (ITAS), Karlsruher Institut für Technologie.

Meyer, Robinson (2015): Could a Bank Deny Your Loan Based on Your Facebook Friends?, in: The Atlantic, Onlineartikel vom 25.09.2015, abrufbar un-

ter: https://www.theatlantic.com/technology/archive/2015/09/facebooks-new-patent-and-digital-redlining/407287/ (letzter Zugriff am 28.8.2019).

Mikians, Jakub; Gyarmati, László; Erramilli, Vijay; Laoutaris, Nikolaos (2012): Detecting price and search discrimination on the internet; in: Proceedings of the 11th ACM Workshop on Hot Topics in Networks, S. 79–84, veröffentlicht von ACM.

Mikians, Jakub; Gyarmati, László; Erramilli, Vijay; Laoutaris, Nikolaos (2013): Crowd-assisted search for price discrimination in e-commerce: First results; in: Proceedings of the ninth ACM conference on Emerging networking experiments and technologies, S. 1–6, veröffentlicht von ACM.

Miller, Akiva A. (2014): What Do We Worry About When We Worry About Price Discrimination? The Law and Ethics of Using Personal Information for Pricing, in: Journal of Technology Law & Policy, 19. Jg., H. 1, S. 41–104.

Milne, George R.; Culnan, Mary J. (2004): Strategies for reducing online privacy risks: Why consumers read (or don't read) online privacy notices, in: Journal of Interactive Marketing, 18. Jg., H. 3, S. 15–29.

Mittelstadt, Brent Daniel; Allo, Patrick; Taddeo, Mariarosaria; Wachter, Sandra; Floridi, Luciano (2016): The ethics of algorithms: Mapping the debate, in: Big Data & Society, 3. Jg., H. 2, S. 1–21.

Moll, Ricarda; Horn, Marco; Scheibel, Lisa; Rusch-Rodosthenous, Miriam (2018): Soziale Medien und die EU-Datenschutzgrundverordnung. Informationspflichten und datenschutzfreundliche Voreinstellungen; Düsseldorf: Marktwächter Digitale Welt, Verbraucherzentrale NRW e.V.

Moos, Flemming; Rothkegel, Tobias (2016): Nutzung von Scoring-Diensten im Online-Versandhandel. Scoring-Verfahren im Spannungsfeld von BDSG, AGG und DS-GVO, in: Zeitschrift für Datenschutz, 6. Jg., H. 12, S. 561–568.

Mullainathan, Sendhil; Spiess, Jann (2017): Machine learning: an applied econometric approach, in: Journal of Economic Perspectives, 31. Jg., H. 2, S. 87–106.

NFHA (2019): National Fair Housing Alliance Settles Lawsuit with Facebook: Transforms Facebook's Ad Platform Impacting Millions of Users,

Onlineartikel veröffentlicht von National Fair Housing Alliance vom 18.03.2019, abrufbar unter: https://nationalfairhousing.org/2019/03/18/national-fair-housing-alliance-settles-lawsuit-with-facebook-transforms-facebooks-ad-platform-impacting-millions-of-users/ (letzter Zugriff am 28.8.2019).

Niklas, Jędrzej (2018): Profiling the Unemployed, in: Digital Society Blog, Humboldt Institut für Internet und Gesellschaft, Onlineartikel vom 16.01.2018, abrufbar unter: https://www.hiig.de/profiling-von-arbeitslosen/ (letzter Zugriff am 28.8.2019).

Niklas, Jędrzej (2019): Polen: Regierung schafft umstrittenes Scoring-System für Arbeitslose ab, Onlineartikel veröffentlicht von AlgorithmWatch vom 16.04.2019, abrufbar unter: https://algorithmwatch.org/story/polnische-regierung-schafft-umstrittenes-scoring-system-fuer-arbeitslose-ab (letzter Zugriff am 28.8.2019).

Niklas, Jędrzej; Sztandar-Sztanderska, Karolina; Szymielewicz, Katarzyna (2015): Profiling the unemployed in Poland: Social and political implications of algorithmic decision making; Warsaw: Fundacja Panoptykon.

Noble, Safiya Umoja (2018): Algorithms of Oppression. How Search Engines Reinforce Racism; New York: New York University Press.

Obermeyer, Ziad; Mullainathan, Sendhil (2019): Dissecting Racial Bias in an Algorithm that Guides Health Decisions for 70 Million People; in: Proceedings of the Conference on Fairness, Accountability, and Transparency, S. 89, veröffentlicht von ACM.

Orwat, Carsten; Bless, Roland (2016): Values and Networks – Steps Toward Exploring their Relationships, in: Computer Communication Review (ACM SIGCOMM), 46. Jg., H. 2, S. 25–31.

Orwat, Carsten; Raabe, Oliver; Buchmann, Erik; Anandasivam, Arun; Freytag, Johan-Christoph; Helberger, Natali; Ishii, Kei; Lutterbeck, Bernd; Neumann, Dirk; Otter, Thomas; Pallas, Frank; Reussner, Ralf; Sester, Peter; Weber, Karsten; Werle, Raymund (2010): Software als Institution und ihre Gestaltbarkeit, in: Informatik-Spektrum, 33. Jg., H. 6, S. 626–633.

Orwat, Carsten; Schankin, Andrea (2018): Attitudes towards big data practices and the institutional framework of privacy and data protection – A po-

pulation survey, KIT Scientific Report 7753; Karlsruhe: KIT Scientific Publishing.

Parasuraman, Raja; Riley, Victor (1997): Humans and automation: Use, misuse, disuse, abuse, in: Human factors, 39. Jg., H. 2, S. 230–253.

Pasquale, Frank (2015): The Black Box Society: the Secret Algorithms that Control Money and Information; Cambridge, London: Harvard University Press.

Penney, Jonathon W. (2016): Chilling Effects: Online Surveillance and Wikipedia Use, in: Berkeley Technology Law Journal, 31. Jg., H. 1, S. 117–182.

Penney, Jonathon W. (2017): Internet surveillance, regulation, and chilling effects online: a comparative case study, in: Internet Policy Review, 6. Jg., H. 2, S. 1–39.

Phelps, Edmund S. (1972): The Statistical Theory of Racism and Sexism, in: The American Economic Review, 62. Jg., H. 4, S. 659–661.

Ponti, Sarah; Tuchtfeld, Erik (2018): Zur Notwendigkeit einer Verbandsklage im AGG, in: Zeitschrift für Rechtspolitik (ZRP), 51. Jg., H. 5, S. 139–141.

Pope, Devin G.; Sydnor, Justin R. (2011): What's in a Picture? Evidence of Discrimination from Prosper.com, in: Journal of Human Resources, 46. Jg., H. 1, S. 53–92.

Powles, Julia; Nissenbaum, Helen (2018): The Seductive Diversion of 'Solving' Bias in Artificial Intelligence, in: Medium, Onlineartikel vom 7.12.2018, abrufbar unter: https://medium.com/s/story/the-seductive-diversion-of-solving-bias-in-artificial-intelligence-890df5e5ef53 (letzter Zugriff am 28.8.2019).

Puri, Ruchir (2018): Mitigating Bias in AI Models, in: IBM Research Blog, Onlineartikel vom 6.2.2018, abrufbar unter: https://www.ibm.com/blogs/research/2018/02/mitigating-bias-ai-models/ (letzter Zugriff am 28.8.2019).

Raabe, Oliver; Wagner, Manuela (2016): Die Zweckbindung: Ein Überblick über die aktuelle Rechtslage und Harmonisierung durch die EU-Datenschutzgrundverordnung, in: Smart-Data-Begleitforschung (Hrsg.): Die Zu-

kunft des Datenschutzes im Kontext von Forschung und Smart Data. Datenschutzgrundprinzipien im Diskurs; Berlin: Smart-Data-Begleitforschung, S. 16–22.

Raabe, Oliver; Wagner, Manuela (2019 in Vorbereitung): Daten, Informationen, Wissen, Entscheiden und Steuern - Ein Referenzmodell für den zukunftsfähigen Datenschutz in wissensbasiert steuernden digitalen Ökosystemen; Karlsruhe: Karlsruher Institut für Technologie, Zentrum für Angewandte Rechtswissenschaft.

Raji, Inioluwa Deborah; Buolamwini, Joy (2019): Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products; in: AAAI/ACM Conf. on AI Ethics and Society.

Reichwald, Julian; Pfisterer, Dennis (2016): Autonomie und Intelligenz im Internet der Dinge. Möglichkeiten und Grenzen autonomer Handlungen, in: Computer und Recht, 32. Jg., H. 3, S. 208–212.

Reidenberg, Joel R. (1998): Lex Informatica: The Formulation of Information Policy Rules Through Technology, in: Texas Law Review, 76. Jg., H. 3, S. 553–584.

Reidenberg, Joel R.; Bhatia, Jaspreet; Breaux, Travis D.; Norton, Thomas B. (2016): Ambiguity in privacy policies and the impact of regulation, in: The Journal of Legal Studies, 45. Jg., H. S2, S. S163–S190.

Reidenberg, Joel R.; Russell, N. Cameron; Callen, Alexander J.; Qasir, Sophia; Norton, Thomas B. (2015): Privacy harms and the effectiveness of the notice and choice framework, in: I/S: A Journal of Law and Policy for the Information Society, 11. Jg., H. 2, S. 485–524.

Reisman, Dillon; Schultz, Jason; Crawford, Kate; Whittaker, Meredith (2018): Algorithmic impact assessments: a practical framework for public agency accountability; New York: AI Now Institute.

Richardson, Rashida; Schultz, Jason; Crawford, Kate (2019): Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice, in: New York University Law Review Online.

Robinson, David; Koepke, Logan (2016): Stuck in a Pattern. Early evidence on “predictive policing” and civil rights; Washington, D.C.: Upturn.

Romei, Andrea; Ruggieri, Salvatore (2014): A multidisciplinary survey on discrimination analysis, in: The Knowledge Engineering Review, 29. Jg., H. 5, S. 582–638.

Romei, Andrea; Ruggieri, Salvatore; Turini, Franco (2013): Discrimination discovery in scientific project evaluation: A case study, in: Expert Systems With Applications, 40. Jg., H. 15, S. 6064–6079.

Rosenblat, Alex; Kneese, Tamara; Boyd, Danah (2014): Networked Employment Discrimination, Open Society Foundations' Future of Work Commissioned Research Papers 2014; New York: Data & Society Research Institute.

Rosenblat, Alex; Levy, Karen E. C.; Barocas, Solon; Hwang, Tim (2017): Discriminating Tastes: Uber's Customer Ratings as Vehicles for Workplace Discrimination, in: Policy & Internet, 9. Jg., H. 3, S. 256–279.

Ru, Hong; Schoar, Antoinette (2016): Do credit card companies screen for behavioral biases?; Washington, D.C.: National Bureau of Economic Research.

Salzburger Nachrichten (2019): Kopf sieht keine Diskriminierung durch AMS-Algorithmus, in: Salzburger Nachrichten, Onlineartikel vom 18.01.2019, abrufbar unter: https://www.sn.at/wirtschaft/oesterreich/kopf-sieht-keine-diskriminierung-durch-ams-algorithmus-64301929 (letzter Zugriff am 28.8.2019).

Sanchez-Monedero, Javier; Dencik, Lina (2018): How to (partially) evaluate automated decision systems; Cardiff: Data Justice Lab.

Sandvig, Christian; Hamilton, Kevin; Karahalios, Karrie; Langbort, Cedric (2014): Auditing algorithms: Research methods for detecting discrimination on internet platforms, Data and discrimination: converting critical concerns into productive inquiry; a preconference at the 64th Annual Meeting of the International Communication Association. May 22, 2014; Seattle, WA, USA.

Schaber, Peter (2012): Menschenwürde; Stuttgart: Reclam.

Schauer, Frederick (2003): Profiles, probabilities, and stereotypes; Cambridge: Harvard University Press.

Schauer, Frederick (2018): Statistical (and non-statistical) discrimination, in: Kasper Lippert-Rasmussen (Hrsg.): The Routledge Handbook of the Ethics of Discrimination; London: Routlegde, S. 42–53.

Scherr, Albert (2016): Diskriminierung/Antidiskriminierung – Begriffe und Grundlagen, in: Aus Politik und Zeitgeschichte, 66. Jg., H. 9, S. 3–10.

Schiek, Dagmar (2000): Differenzierte Gerechtigkeit: Diskriminierungsschutz und Vertragsrecht; Baden-Baden: Nomos.

Schinzel, Britta (2017): Algorithmen sind nicht schuld, aber wer oder was ist es dann?, in: FIfF-Kommunikation, H. 2, S. 5–9.

Schneider, Ingrid; Ulbricht, Lena (2018): Ist Big Data fair? Normativ hergestellte Erwartungen an Big Data, in: Barbara Kolany-Raiser, Reinhard Heil, Carsten Orwat und Thomas Hoeren (Hrsg.): Big Data und Gesellschaft. Eine multidisziplinäre Annäherung; Wiesbaden: Springer VS, S. 198–207.

Scholz, Philip (2019): DSGVO Art. 22 Automatisierte Entscheidungen im Einzelfall einschließlich Profiling, in: Spiros Simitis, Gerrit Hornung und Indra Spiecker genannt Döhmann (Hrsg.): Datenschutzrecht. DSGVO und BDSG; Baden-Baden: Nomos.

Schrader, Peter; Schubert, Jens (2018): AGG §3 Begriffsbestimmungen, in: Wolfgang Däubler und Martin Bertzbach (Hrsg.): Allgemeines Gleichbehandlungsgesetz – Handkommentar, 4. Aufl.; Baden-Baden: Nomos.

Schwaiger, Manfred; Hufnagel, Gerrit (2018): Handel und elektronische Bezahlsysteme, Gutachten im Rahmen des Projekts „Assessing Big Data“ (ABIDA); München: Ludwig-Maximilians-Universität München, Institut für Marktorientierte Unternehmensführung.

Schwartz, Paul M. (1999): Privacy and democracy in cyberspace, in: Vanderbilt Law Review, 52. Jg., H. 6, S. 1607–1702.

Schweighofer, Erich; Sorge, Christoph; Borges, Georg; Schäfer, Burkhard; Waltl, Bernhard; Grabmair, Matthias; Krupka, Daniel (2018): Technische und rechtliche Betrachtungen algorithmischer Entscheidungsverfahren. Gutachten der Fachgruppe Rechtsinformatik der Gesellschaft für Informatik e.V. im Auftrag des Sachverständigenrats für Verbraucherfragen; Berlin:

Sachverständigenrat für Verbraucherfragen (SVRV) beim Bundesministerium der Justiz und für Verbraucherschutz.

Selbst, Andrew D. (2017): Disparate Impact in Big Data Policing, in: Georgia Law Review, 52. Jg., H. 1, S. 109–195.

Selbst, Andrew D.; Barocas, Solon (2018): The Intuitive Appeal of Explainable Machines, in: Fordham Law Review, 87. Jg., H. 3, S. 1085–1139.

Selke, Stefan; Biniok, Peter; Achatz, Johannes; Späth, Elisabeth (2018): Ethische Standards für Big Data und deren Begründung, Gutachten im Rahmen des Projekts „Assessing Big Data" (ABIDA); Münster, Karlsruhe: Universität Münster, Karlsruher Institut für Technologie.

Shah, Rajiv C.; Kesan, Jay P. (2010): Software as Governance, in: Hans J. Scholl (Hrsg.): E-Government Information, Technology, and Transformation; Armonk: M.E. Sharpe, S. 125–140.

Sherwin, Galen; Bhandari, Esha (2019): Facebook Settles Civil Rights Cases by Making Sweeping Changes to Its Online Ad Platform, Onlineartikel veröffentlicht von American Civil Liberties Union (ACLU) vom 19.03.2019, abrufbar unter: https://www.aclu.org/blog/womens-rights/womens-rights-workplace/facebook-settles-civil-rights-cases-making-sweeping (letzter Zugriff am 28.8.2019).

Silver, Joe (2013): Is Your Turn-By-Turn Navigation Application Racist?, Onlineartikel veröffentlicht von American Civil Liberties Union (ACLU vom 2.10.2013), abrufbar unter: https://www.aclu.org/blog/national-security/your-turn-turn-navigation-application-racist (letzter Zugriff am 27.8.2019).

Smeddinck, Ulrich; Bornemann, Basil (2018): Verkehr, Mobilität, Nudging - Zugleich zum Stand von Regulieren durch Anstoßen in Deutschland, in: Die Öffentliche Verwaltung, 71. Jg., H. 13, S. 513–523.

Snow, Jacob (2018): Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots, Onlineartikel veröffentlicht von American Civil Liberties Union (ACLU) vom 26.07.2018, abrufbar unter: https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28 (letzter Zugriff am 28.8.2019).

Solove, Daniel J. (2013): Privacy Self-Management and the Consent Dilemma, in: Harvard Law Review, 126. Jg., H. 7, S. 1880–1903.

Spielkamp, Mathias (Hrsg.) (2019): Automating Society. Taking Stock of Automated Decision-Making in the EU. A report by AlgorithmWatch in cooperation with Bertelsmann Stiftung, supported by the Open Society Foundations; Berlin: AW AlgorithmWatch gGmbH.

Starr, Sonja B. (2014): Evidence-based sentencing and the scientific rationalization of discrimination, in: Stanford Law Review, 66. Jg., H. 4, S. 803–872.

Steppe, Richard (2017): Online price discrimination and personal data: A General Data Protection Regulation perspective, in: Computer Law & Security Review, 33. Jg., H. 6, S. 768–785.

Straker, Christian; Niehoff, Maurice (2018): ABIDA-Fokusgruppe – Diskriminierung durch Algorithmen und KI im eRecruiting, in: ZD-Aktuell, H. 06252.

Strauß, Stefan (2018): From Big Data to Deep Learning: A Leap Towards Strong AI or 'Intelligentia Obscura'?, in: Big Data and Cognitive Computing, 2. Jg., H. 3, S. 1–19.

Sunstein, Cass R. (2014): Nudging: A Very Short Guide, in: Journal of Consumer Policy, 37. Jg., H. 4, S. 583–588.

Supik, Linda (2017): Statistik und Diskriminierung, in: Albert Scherr, Aladin El-Mafaalani und Gökçen Yüksel (Hrsg.): Handbuch Diskriminierung; Wiesbaden: Springer VS, S. 191–207.

SVRV (2018): Verbrauchergerechtes Scoring. Gutachten des Sachverständigenrats für Verbaucherfragen; Berlin: Sachverständigenrat für Verbaucherfragen (SVRV).

Swedloff, Rick (2014): Risk classification's big data (r) evolution, in: Connecticut Insurance Law Journal, 21. Jg., H. 1, S. 339–373.

Sweeney, Latanya (2013): Discrimination in Online Ad Delivery, in: Communication of the ACM, 56. Jg., H. 5, S. 44–54.

Szigetvari, András (2018): Beruf, Ausbildung, Alter, Geschlecht: Der Algorithmus des AMS, in: DerStandard, Onlineartikel vom 15.10.2018, abrufbar unter: https://derstandard.at/2000089325546/Beruf-Ausbildung-Alter-Geschlecht-Das-sind-die-Zutaten-zum-neuen (letzter Zugriff am 28.8.2019).

Tatman, Rachael (2017): Gender and dialect bias in YouTube's automatic captions; in: Proceedings of the First ACL Workshop on Ethics in Natural Language Processing, S. 53–59.

ten Have, Marieke (2013): Lifestyle Differentiation in Health Insurance. An Overview of the Ethical Arguments, Monitoring Report on Ethics and Health 2013; The Hague: Netherlands Centre for Ethics and Health.

The Royal Society (2017): Machine learning: the power and promise of computers that learn by example; London: The Royal Society.

The White House (2014): Big Data: Seizing Opportunities, Preserving Values; Washington, D.C.: The White House, Executive Office of the President.

Thelwall, Mike (2018): Gender bias in sentiment analysis, in: Online Information Review, 42. Jg., H. 1, S. 45–57.

Tillmann, Tristan Julian; Vogt, Verena (2018a): Personalisierte Preise – Diskriminierung 2.0?, ABIDA-Dossier; Münster, Karlsruhe: Projekt „Assessing Big Data" (ABIDA).

Tillmann, Tristan Julian; Vogt, Verena (2018b): Personalisierte Preise im Big-Data-Zeitalter, in: Verbraucher und Recht (VuR), 33. Jg., H. 12, S. 447–455.

Tolan, Songül (2018): Fair and Unbiased Algorithmic Decision Making: Current State and Future Challenges; Seville: European Commission, Joint Research Centre.

Tolan, Songül; Miron, Marius; Gómez, Emilia; Castillo, Carlos (2019): Why Machine Learning May Lead to Unfairness: Evidence from Risk Assessment for Juvenile Justice in Catalonia; in: ICAIL '19, June 17–21, 2019, Montreal, QC, Canada.

Tramèr, Florian; Atlidakis, Vaggelis; Geambasu, Roxana; Hsu, Daniel; Hubaux, Jean-Pierre; Humbert, Mathias; Juels, Ari; Lin, Huang (2017): FairTest: Discovering unwarranted associations in data-driven applications; in: 2017 IEEE European Symposium on Security and Privacy (EuroS&P), S. 401–416, veröffentlicht von IEEE.

Trute, Hans-Heinrich (1998): Der Schutz personenbezogener Informationen in der Informationsgesellschaft, in: Juristenzeitung, 53. Jg., H. 17, S. 822–831.

Trute, Hans-Heinrich (2003): Verfassungsrechtliche Grundlagen, in: Alexander Roßnagel (Hrsg.): Handbuch Datenschutzrecht; München: C.H. Beck, S. 156–187.

ULD; GP Forschungsgruppe (2014): Scoring nach der Datenschutz-Novelle 2009 und neue Entwicklungen. Abschlussbericht; Kiel, München: Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein (ULD); GP Forschungsgruppe.

UN GA (2018): Promotion and protection of the right to freedom of opinion and expression. Seventy-third session, 29 August 2018; New York: United Nations, General Assembly (UN GA).

UNESCO (1951): Statement on Race; Paris: United Nations Educational, Scientific and Cultural Organization (UNESCO).

United States District Court for the Northern District of California (2018): Communications Workers of America v. T-Mobile US Inc, First Amended Class and Collective Action Complaint – Demand for Jury Trial. Case No. 17-cv-07232-BLF, Onlineartikel, abrufbar unter: https://www.onlineage-discrimination.com/sites/default/files/documents/og-cwa-complaint.pdf (letzter Zugriff am 28.8.2019).

US CEA (2015): Big Data and Differential Pricing; Washington, D.C.: Council of Economic Advisers (CEA), Executive Office of the President of the United States.

US HUD (2019a): Charge of Discrimination. HUD versus Facebook; Washington, D.C.: United States of America, Department of Housing and Urban Development, Office of Administrative Law Judges.

US HUD (2019b): HUD charges Facebook with housing discrimination over company's targeted advertising practices, Onlineartikel veröffentlicht von U.S. Department of Housing and Urban Development (HUD) vom 28.03.2019, abrufbar unter: https://www.hud.gov/press/press_releases_media_advisories/HUD_No_19_035 (letzter Zugriff am 28.8.2019).

Van Alsenoy, Brendan; Kosta, Eleni; Dumortier, Jos (2014): Privacy notices versus informational self-determination: Minding the gap, in: International Review of Law, Computers and Technology, 28. Jg., H. 2, S. 185–203.

Van Otterlo, Martijn (2013): A machine learning view on profiling, in: Mireille Hildebrandt und Katja de Vries (Hrsg.): Privacy, Due Process and the Computational Turn. The philosophy of law meets the philosophy of technology; Abingdon: Routledge, S. 67–90.

Varian, Hal R. (2014): Beyond Big Data, in: Business Economics, 49. Jg., H. 1, S. 27–31.

Varian, Hal R.; Farrell, Joseph; Shapiro, Carl (2004): The Economics of Information Technology: An Introduction; Cambridge et al.: Cambridge University Press.

Vercellis, Carlo (2011): Business Intelligence: Data Mining and Optimization for Decision Making; Chichester: Wiley.

Verma, Sahil; Rubin, Julia (2018): Fairness definitions explained; in: 2018 IEEE/ACM International Workshop on Software Fairness (FairWare), S. 1–7, veröffentlicht von IEEE.

Volkova, Svitlana; Bachrach, Yoram (2015): On predicting sociodemographic traits and emotions from communications in social networks and their implications to online self-disclosure, in: Cyberpsychology, Behavior, and Social Networking, 18. Jg., H. 12, S. 726–736.

von Grafenstein, Max; Hölzel, Julian; Irgmaier, Florian; Pohle, Jörg (2018): Nudging - Regulierung durch Big Data und Verhaltenswissenschaften, Gutachten im Rahmen des Projekts „Assessing Big Data" (ABIDA); Münster, Karlsruhe: Universität Münster, Karlsruher Institut für Technologie.

Wachter, Sandra; Mittelstadt, Brent; Floridi, Luciano (2017): Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General

Data Protection Regulation, in: International Data Privacy Law, 7. Jg., H. 2, S. 76–99.

Wang, Yilun; Kosinski, Michal (2018): Deep neural networks are more accurate than humans at detecting sexual orientation from facial images, in: Journal of Personality and Social Psychology, 114. Jg., H. 2, S. 246–257.

Wei, Yanhao; Yildirim, Pinar; Van den Bulte, Christophe; Dellarocas, Chrysanthos (2016): Credit Scoring with Social Network Data, in: Marketing Science, 35. Jg., H. 2, S. 234–258.

Weichert, Thilo (2013): Big Data und Datenschutz - Chancen und Risiken einer neuen Form der Datenanalyse, in: Zeitschrift für Datenschutz, 3. Jg., H. 6, S. 251–259.

Weichert, Thilo (2014): Scoring in Zeiten von Big Data, in: Zeitschrift für Rechtspolitik (ZRP), 47. Jg., H. 6, S. 168–171.

Weichert, Thilo (2018): Big Data im Gesundheitsbereich, Gutachten im Rahmen des Projekts „Assessing Big Data“ (ABIDA); Münster, Karlsruhe: Universität Münster; Karlsruher Institut für Technologie.

Wersig, Maria (2017): Fälle zum Allgemeinen Gleichbehandlungsgesetz (AGG). Eine Einführung in Theorie und Praxis des Antidiskriminierungsrechts in 23 Fällen; Opladen, Toronto: Verlag Barbara Budrich.

Wiegerling, Klaus (2016): Würde, Autonomie, Subsidiarität – Ist das alles? Ist das viel?; in: Die Werte des Westens. Wofür wir stehen und werben sollten, Vortrag an der Hochschule Konstanz, 30.05.2016.

Wiegerling, Klaus; Nerurkar, Michael; Wadephul, Christian (2018): Ethische und anthropologische Aspekte der Anwendung von Big-Data-Technologien, in: Barbara Kolany-Raiser, Reinhard Heil, Carsten Orwat und Thomas Hoeren (Hrsg.): Big Data und Gesellschaft. Eine multidisziplinäre Annäherung; Wiesbaden: Springer VS, S. 1–67.

Williams, Betsy Anne; Brooks, Catherine F.; Shmargad, Yotam (2018): How algorithms discriminate based on data they lack: Challenges, solutions, and policy implications, in: Journal of Information Policy, 8. Jg., S. 78–115.

Wilson, Benjamin; Hoffman, Judy; Morgenstern, Jamie (2019): Predictive Inequity in Object Detection, in: arXiv preprint arXiv:1902.11097.

Wimmer, Barbara (2018a): AMS-Chef: „Mitarbeiter schätzen Jobchancen pessimistischer ein als der Algorithmus". Interview mit Johannes Kopf, Vorstand des Arbeitsmarktservice (AMS) am 21.10.2015 in Wien, in: futurezone, Onlineartikel vom 12.10.2018, abrufbar unter: https://futurezone.at/netzpolitik/ams-chef-mitarbeiter-schaetzen-jobchancen-pessimistischer-ein-als-der-algorithmus/400143839 (letzter Zugriff am 28.8.2019).

Wimmer, Barbara (2018b): „AMS-Sachbearbeiter erkennen nicht, wann ein Programm falsch liegt", in: futurzone, Onlineartikel vom 18.10.2018, abrufbar unter: https://futurezone.at/netzpolitik/ams-sachbearbeiter-erkennen-nicht-wann-ein-programm-falsch-liegt/400147472 (letzter Zugriff am 28.8.2019).

WIPO (2019): WIPO Technology Trends 2019 - Artificial Intelligence; Genf: World Intellectual Property Organization (WIPO).

Wischmeyer, Thomas (2018): Regulierung intelligenter Systeme, in: Archiv des öffentlichen Rechts, 143. Jg., H. 1, S. 1–66.

World Wide Web Foundation (2017): Algorithmic Accountability. Applying the Concept to Different Country Contexts; Washington, D.C.: World Wide Web Foundation.

WP29 (2017a): Guidelines on Consent under Regulation 2016/679; Brussels: Article 29 Data Protection Working Party (WP29).

WP29 (2017b): Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679; Brussels: Article 29 Data Protection Working Party (WP29).

Wu, Xiaolin; Zhang, Xi (2016): Automated inference on criminality using face images, in: arXiv preprint arXiv:1611.04135, S. 4038–4052.

Yeung, Karen (2008): Towards an Understanding of Regulation by Design, in: Roger Brownsword und Karen Yeung (Hrsg.): Regulating Technologies. Legal Futures, Regulatory Frames and Technological Fixes; Oxford and Portland: Hart, S. 79–107.

Yeung, Karen (2017): Algorithmic regulation: A critical interrogation, in: Regulation & Governance, 12. Jg., H. 4, S. 505–523.

Yeung, Karen (2018): A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework; Strasbourg: Council of Europe, Committee of experts on human rights dimensions of automated data processing and different forms of artificial intelligence (MSI-AUT).

Yue, Lin; Chen, Weitong; Li, Xue; Zuo, Wanli; Yin, Minghao (2018): A survey of sentiment analysis in social media, in: Knowledge and Information Systems, 60. Jg., H. 2, S. 617–663.

YVTltk (2018): Assessment of creditworthiness, authority, direct multiple discrimination, gender, language, age, place of residence, financial reasons, conditional fine. Plenary Session (voting), Register number: 216/2017, 21 March 2018; Finland, Government Publication: Yhdenvertaisuus- ja tasa-arvolautakunta/National Non-Discrimination and Equality Tribunal of Finland.

Zander-Hayat, Helga; Reisch, Lucia A.; Steffen, Christine (2016): Personalisierte Preise: Eine verbraucherpolitische Einordnung, in: Verbraucher und Recht, 31. Jg., H. 11, S. 403–409.

Zarsky, Tal Z. (2016): The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making, in: Science, Technology & Human Values, 41. Jg., H. 1, S. 118–132.

Žliobaitė, Indrė; Custers, Bart (2016): Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models, in: Artificial Intelligence and Law, 24. Jg., H. 2, S. 183–201.

Zuboff, Shoshana (2015): Big other: surveillance capitalism and the prospects of an information civilization, in: Journal of Information Technology, 30. Jg., H. 1, S. 75–89.

Zuiderveen Borgesius, Frederik (2016): Singling out people without knowing their names - Behavioural targeting, pseudonymous data, and the new Data Protection Regulation, in: Computer Law and Security Review, 32. Jg., H. 2, S. 256–271.

Zuiderveen Borgesius, Frederik (2018): Discrimination, artificial intelligence, and algorithmic decision-making; Strasbourg: Council of Europe, Directorate General of Democracy.

Zuiderveen Borgesius, Frederik; Poort, Joost (2017): Online Price Discrimination and EU Data Privacy Law, in: Journal of Consumer Policy, 40. Jg., H. 3, S. 347–366.

Zweig, Katharina (2019): Algorithmische Entscheidungen: Transparenz und Kontrolle, Bericht Analysen und Argumente. Digitale Gesellschaft, Nr. 338; Sankt Augustin: Konrad Adenauer Stiftung.

Zweig, Katharina; Fischer, Sarah; Lischka, Konrad (2018): Wo Maschinen irren können. Fehlerquellen und Verantwortlichkeiten in Prozessen algorithmischer Entscheidungsfindung; Gütersloh: Bertelsmann Stiftung.

Impressum

Diese Publikation ist Teil der Öffentlichkeitsarbeit der
Antidiskriminierungsstelle des Bundes;
sie wird kostenlos abgegeben und ist nicht zum Verkauf bestimmt.

Herausgeberin:
Antidiskriminierungsstelle des Bundes
11018 Berlin
www.antidiskriminierungsstelle.de

Autor:
Dr. Carsten Orwat

Kontakt:
Tel.: +49(0) 30 18555-1855
Fax: +49(0) 30 18555-41865
Juristische Erstberatung: Mo. 13–15 Uhr, Mi. und Fr., 9–12 Uhr
E-Mail: **beratung@ads.bund.de**
Allgemeine Anfragen: Mo. bis Fr., 9–12 Uhr und 13–15 Uhr
E-Mail: **poststelle@ads.bund.de**

Gestaltung: www.zweiband.de
Stand: September 2019, 1. Auflage
Druck: MKL Druck GmbH & Co. KG

ISBN 978-3-8487-6285-9

Algorithmen, unter anderem der künstlichen Intelligenz, werden in vielfältiger Weise für Differenzierungen von Personen, Diensten, Produkten, Positionen oder beim staatlichen Handeln eingesetzt. Die vorliegende Studie zeigt anhand von Beispielfällen nicht nur technische und organisatorische Ursachen von Diskriminierungsmöglichkeiten, sondern vor allem auch die gesellschaftlichen Risiken auf. Sie rufen einen Bedarf nach Reformen des Antidiskriminierungs- und Datenschutzrechts hervor, aber ebenso gesellschaftliche Abwägungen und Festlegungen, welche algorithmen- und datenbasierten Differenzierungen in einer Gesellschaft überhaupt für akzeptabel gehalten werden. Nicht zuletzt werden Aufgaben für Antidiskriminierungsstellen diskutiert, die von der Identifizierung und dem Nachweis von algorithmenbasierten Diskriminierungen bis hin zu präventivem und kooperativem Vorgehen reichen.